COMPUTATIONAL ANALYSIS OF CIS-REGULATORY MECHANISMS UNDERLYING CELLULAR RESPONSE TO DRUG PERTURBATIONS

BY

CASEY HANSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Saurabh Sinha, Chair
Dr. Liewei Wang, Mayo Clinic Rochester
Associate Professor Jian Peng
Professor Ravishankar K. Iyer

ABSTRACT

In this thesis, I, alongside my varied and talented collaborators, aim to integrate two seemingly disparate subfields of biology: pharmacogenomics and regulatory genomics. The former concentrates on how inherited factors such as SNPs and genes inform the response to medical interventions designed to ameliorate disease states, processes, and phenotypes; ontologically, it resides at the intersection of genomics and pharmacology. Regulatory genomics, however, directs its attention fully to the biological mechanisms by which genes express themselves, without regard to application or context. As such, it is viewed within the intersection of genomics and systems biology. A systems perspective is relevant because last 20 years of regulatory genomics research have concluded that (a) genes do not necessarily express binarily as a (b) function of logical operations on the presence/absence of a (c) finite set of protein regulators binding to the (d) genes immediate upstream DNA sequence, but rather, that the span of regulators ranges from protein DNA binding factors to other such nuclear factors (including the DNA itself) up and through various signaling molecule networks to the various components of the cell, terminating at the interface of the cell and extracellular environment the membrane. DNA accessibility and conformation, cellular type, micro-RNAs are all real factors that influence gene expression. This thesis, however, assumes the more nave assumptions of gene regulation namely that the DNA binding proteins, called Transcription Factors (TFs), influence the kinetics of the transcriptional process by binding in certain configurations proximal to the genes transcription start site. In combining both pharmacogenomics and regulatory genomics, it was necessary to make simplifications and assumptions to the biology on both sides to make the problem tractable; however, as a result, we were able to construct statistical and probabilistic models that capture novel biological associations that would otherwise remain unknown. Namely, associations between nuclear regulators of genes and drug response. In doing so, we provide a novel link in connecting these two disparate views of biology into a more cohesive and singular picture.

To my parents, for their love and support.

ACKNOWLEDGMENTS

Funding for this work was provided in part by the National Institute of Health (NIH) (R01 GM114341 to SS, Grant U54 GM114838 to SS, U19 GM61388 Pharmacogenomics Research Network and R01 CA138461 to LW) and in part by the Mayo Clinic-University of Illinois at Urabana-Champaign (UIUC) Alliance and by Grant 1U54GM114838 awarded by the National Institute of General Medical Sciences (NIGMS) through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

I would like to thank Dr. Liewei Wang, Dr. Junmei Cairns, and Dr. Richard Weinshilboum for their continued advice and support from collaborating with Dr. Sinha and I on the two critical papers of this work to hosting me at the Mayo Clinic. I would also like to thank all former and current Sinha Lab members for their help and encouragement over the years. Of course, I would like to thank my advisor Dr. Saurabh Sinha for supporting me in so many ways throughout this doctoral program. Its been a long journey. And finally, I would like to thank all of the committee members not yet named, Dr. Jian Peng and Dr. Ravishankar Iyar, for their insight on my prelim committee and feedback on my defense and this current work which have proven invaluable in constructing this thesis.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The field of pharmacogenetics aims to understand the relationship between inter-individual variation at the genetic level and variation in cellular or physiological response to drug administration. Rapidly emerging genomic technologies have expanded the scope of analysis to genome-wide levels, simultaneously providing a variety of high quality data to enable the analysis, including genotype[1] , gene[2] expression[3] and proteomic data[4] , as well as functional annotations from the Encyclopedia of Genomic Elements (ENCODE) [1]. Consequently, the burden has fallen to computational analysts to integrate and explain the current deluge of biological data, in a manner consistent with both the underlying biological principles and mechanisms, while also delivering novel insights into the molecular mechanisms underlying cellular response to drug treatment. This thesis effort to develop statistical and probabilistic models to elucidate the relationship between regulatory factors and cellular perturbagens. It is divided into two such chapters: Chapter 2 corresponding to a simple statistical model that integrates much of our intuition regarding the relationship between primitive biological entities and Chapter 3 corresponding to a re-imagining of the model in Chapter 2 in a probabilistic sense so as to increase model extensibility, interpretability, and interrogability. To clarify, in shifting from a statistical enrichment test to a probabilistic graphical model (PGM) that extends the statistical model in a meaningful way, allowing for greater interpretability and integration of heterogeneous data types. We introduce each of these studies in the following sub-sections.

1.1 GENE EXPRESSION IN THE MIDDLE (GENMI)

This study integrates gene expression, genotype and drug response data in lymphoblastoid cell lines with transcription factor (TF)-binding sites from ENCODE (Encyclopedia of Genomic Elements) in a novel methodology that elucidates regulatory contexts associated with cytotoxicity. The method, GENMi (Gene Expression iN the Middle), postulates that

[1]The specific sequence of DNA units (nucleotides or base pairs) which constitute the genome of an organism. A specific nucleotide can assume one of four types: Adenine, Cytosine, Guanine, and Thymine abbreviated A,C,G,T. The scale at which genotype refers can vary dramatically in context: from the entire genome, to DNA segments, to single base pairs.

[2]A type of DNA segment whose composition encodes recipes to make one or more types of proteins.

[3]The product or process of transcription, whereby a genes DNA is transcribed into a molecule, mRNA, that can move freely toward cellular components necessary for protein production.

[4]Measurements on the quantity and concentration of proteins in a sample. The proteome refers to all such proteins in a genome while transcriptome refers to all such genes. The gene analog to proteomics is transcriptomics: the measurements on the quantity and concentration of mRNA transcribed from genes.

single-nucleotide polymorphisms within TF-binding sites putatively modulate its regulatory activity, and the resulting variation in gene expression leads to variation in drug response. Analysis of 161 TFs and 24 treatments revealed 334 significantly associated TFtreatment pairs. Investigation of 20 selected pairs yielded literature support for 13 of these associations, often from studies where perturbation of the TF expression changes drug response. Experimental validation of significant GENMi associations in taxanes and anthracyclines across two triple-negative breast cancer cell lines corroborates our findings. The method is shown to be more sensitive than an alternative, genome-wide association study-based approach that does not use gene expression. These results demonstrate the utility of GENMi in identifying TFs that influence drug response and provide a number of candidates for further testing. Advances in the burgeoning field of pharmacogenomics [2] have the potential to revolutionize health care by guiding personalized health care for patients via genome sequencing [3, 4].

1.1.1 BACKGROUND

The de facto method for generating biological hypotheses of clinical relevance involves the extraction of pharamacogenomic data from human cell lines that are generalizable and easily manipulated; lymphoblastoid cell lines (LCLs) represent a canonical source with such clear clinical utility [4, 5, 6]. A number of studies have analyzed such data sets to relate genotypic variation to drug response [7], revealing important single-nucleotide polymorphisms (SNPs) as well as SNP-carrying genes that are candidates for functional testing. Parallel to the identification of SNPs associated with drug response, there is also considerable interest in characterizing the mechanistic basis of such relationships, that is, pathways and regulatory networks involved in drug response and its variability [8, 9].

Identifying systems-level components of the response, such as signaling pathways and transcriptional cascades, can enable discovery of novel drug targets and lead to the realization of precision medicine [10]. Here, we embark upon one such line of enquiryto identify transcription factors (TFs) whose regulatory activities are associated with cellular response to cytotoxic treatments (Fig 1.1A), with the expectation that in the future the response may be manipulated by intervening with the function of TF.

The most widely used statistical method for harnessing pharamacogenomic data to identify biomarkers relevant to drug-induced cellular response is genome-wide association study (GWAS), where SNPs are analyzed for their correlation with drug response across individuals. The multigenic origins of phenotypic variability, correlations between proximally

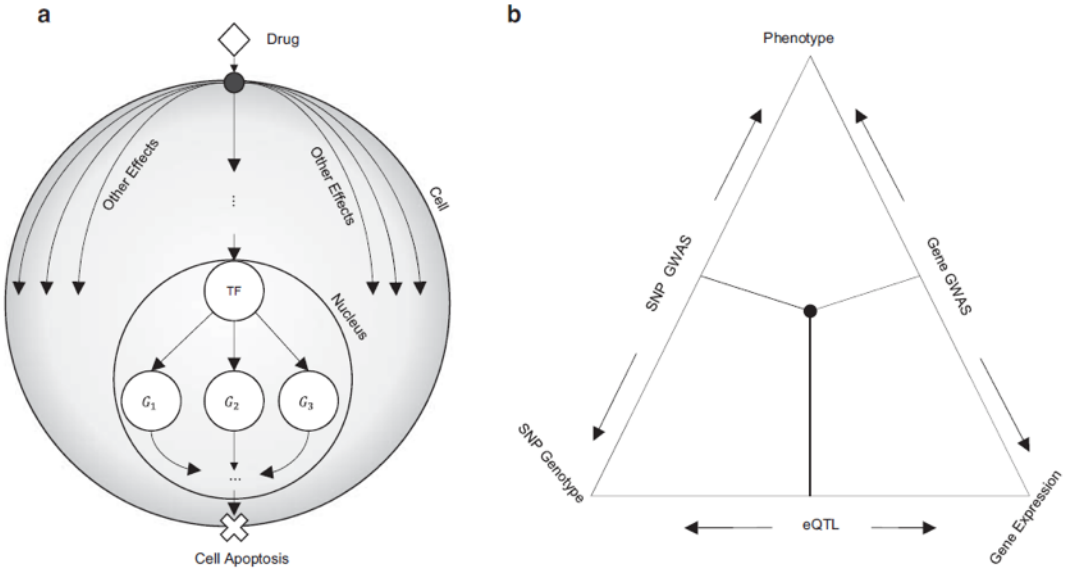


Figure 1.1: (a) Diagram of how transcription factors (TFs) mediate response to a drug. A drug (diamond) enters the cell and affects multiple cellular processes. One such process involves transport or signal transduction to the nucleus where it alters the transcriptional activity of a TF. Expression of target genes is subsequently altered, potentially resulting in apoptosis. (b) Outline of triangulation procedure proposed in the literature. Each edge of the triangle corresponds to a correlation between two of the three axes of information: drug response, genetic variants and gene expression. Integrative analysis involves intersecting expression quantitative trait locus (eQTL) genes and genome-wide association study (GWAS) genes or eQTL single-nucleotide polymorphisms (SNPs) and GWAS SNPs

located SNPs and the need for multiple hypothesis correction over millions of candidate markers reduce the ability of GWAS to discover casual SNPs. A number of studies have sought to improve upon the basic GWAS approach, for example, by testing subsets of SNPs as opposed to single marker analysis, or combining GWAS associations with prior knowledge of gene networks and pathways [11, 12, 13]. We draw inspiration from this emerging paradigm and attempt to associate drug response variation with multiple SNPs that share a common functional context, viz., that of being located within binding sites of the same TF.

Functional genotypic variants are expected to exert their influence on phenotypic differences at least partly through variation in expression levels of nearby genes [14, 15, 16]. A previous study [17] argued that if gene expression data are available in addition to genotype and phenotype data on the same cohort, then a statistical approach called mediator analysis can be employed to discover functional SNPs with greater sensitivity. Gene expression and proteomic data have often been used to predict phenotypes, including drug-induced cellular

response [16, 18, 19, 20, 21]. Efforts have been made to develop methodologies that integrate both gene expression and genotype information to predict phenotype. One method of achieving this is the following: first conduct a GWAS associating SNPs with the phenotype, and then correlate significant GWAS SNPs with expression of their proximal genes, thereby identifying expression quantitative trait locus (eQTL) SNPs, and finally correlate expression of these eQTL genes with phenotype. This triangulation procedure for integrating genotype, gene expression and phenotype (Fig 1.1B) data has been used to identify either candidate SNPs or candidate genes for experimentation. It has also been used in pharmacogenomics to identify biomarkers and genes related to cisplatin-, etoposide- and radiation-induced cytotoxicity [22, 23, 24]. These previous studies motivate us to integrate gene expression data in our search for molecular determinants of drug response variability.

Integration of genotype and expression data in association studies draws our attention to cis-regulatory SNPs that represent a large proportion of individual variability and have been linked to important phenotypes, including diseases [25]. A prime difficulty in characterizing regulatory variants is the poor annotation of the noncoding genome, making it difficult to tell neutral from potentially functional cis-variants. Recent community efforts to systematically annotate the regulatory genome, such as the ENCODE project, may alleviate this problem to an extent. However, few existing approaches incorporate ENCODE data, in particular, transcription factor-binding site (TFBS) data, into statistical analysis of individual variation at the genotypic and phenotypic levels. The ENCODE consortium analyzed the overlap of disease-associated SNPs from the NHGRI GWAS catalog with TFBS and DNase I hypersensitivity sites [1]. A recent study integrated ENCODE data, among other sources of functional data, into a model that selects optimally informative annotation filters to improve SNP association studies [26]. Another study concluded that GWAS SNPs embedded in cis-regulatory elements from disease-relevant cell types are likely to function as eQTLs [27]. However, these studies do not provide a systematic method for integrating all of the above-mentioned types of genomic data so as to determine candidate regulators of phenotypic variation. Our study aims to address this issue by developing a computational method named GENMi (Gene Expression iN the Middle) that integrates ENCODE TFBS, genotype, gene expression and drug induced cytotoxicity data in LCLs to quantify the association between a TF and drug response, thereby elucidating putative regulators responsible for observed cellular responses to drugs.

1.2 PROBABILISTIC GENE EXPRESSION IN THE MIDDLE (PGENMI)

In this study, we develop a new probabilistic model, called pGENMi, that integrates multi-omic data to investigate the transcriptional regulatory mechanisms underlying inter-individual variation of a specific phenotype - that of cell line response to cytotoxic treatment. In particular, pGENMi simultaneously analyzes genotype, DNA methylation, gene expression and transcription factor (TF)DNA binding data, along with phenotypic measurements, to identify TFs regulating the phenotype. It does so by combining statistical information about expression quantitative trait loci (eQTLs) and expression-correlated methylation marks (eQTMs) located within TF binding sites, as well as observed correlations between gene expression and phenotype variation. Application of pGENMi to data from a panel of lymphoblastoid cell lines treated with 24 drugs, in conjunction with ENCODE TF ChIP data, yielded several known as well as novel (TF, Drug) associations. Experimental validations by TF knock-down confirmed 41% of the predicted and tested associations, compared to a 12% confirmation rate of tested non-associations (controls). Extensive literature survey also corroborated 62% of the predicted associations above a stringent threshold. Moreover, associations predicted only when combining eQTL and eQTM data showed higher precision compared to an eQTL-only or eQTM-only analysis using pGENMi, further demonstrating the value of multi-omic integrative analysis.

1.2.1 BACKGROUND

There is great interest today in understanding why certain drugs are effective in some individuals but less so in others. Many studies have sought to identify mechanisms of action of specific drugs [28, 29, 30] as well as genotypic variations that are predictive of an individuals drug response [31, 4, 32]. A major class of drugs of interest today are cytotoxic drugs that may be used in cancer treatment. Large scale data generation efforts, including genotypic and molecular profiling of panels of cell lines [33, 34] along with drug response (cytotoxicity) measurement on those cell lines [33, 35, 36], are expected to facilitate future advances in cancer pharmacogenomics. As the diversity of such data sets increases, it is important to devise rigorous computational methods that can combine these diverse data in a principled manner to help scientists answer mechanistic as well as therapeutic questions pertaining to drug response. For instance, correlating gene expression and drug response in a panel of cell lines helps identify cytotoxicity-related genes [36] but it is not clear how one might extend the approach to additionally exploit genotype (SNP) and epigenotype data (e.g., CpG methylation marks) to maximum effect. We need a statistically sound approach

capable of modeling phenotypic variation while integrating several heterogeneous genomic and epigenomic data types. Additionally, despite numerous efforts to connect phenotype with genotype and regulatory elements, there has not been a systematic effort to aggregate such connections to learn major regulatory mechanisms underlying the genotype-phenotype relationship and its variation across individuals. Furthermore, the regulatory impact of epigenetic sources of variation (for instance CpG methylation) are usually assessed in isolation of genetic variants (SNPs). Recent studies have shown that there may be a complex interplay between genetic, epigenetic, and transcriptional variation in relation to disease [37], arguing for a more integrative approach to their analysis.

We present here a novel, statistically principled approach to aggregating data on genetic as well as epigenetic variations, along with genome-wide profiles of regulatory function, to derive associations between a transcription factor and individual variation in cytotoxic response to a drug; this permits a mechanistic interpretation of the impact of molecular variants on drug response. There have been new insights into how TFs may be regulated by small molecules [38, 39], and there is no doubt as to their significance in human diseases such as cancer [40, 41, 42]. Our aim therefore is to push the frontier of knowledge regarding the relationship of TFs and drug response for putative therapeutic benefit. Our new computational method is based on a statistical formalism called probabilistic graphical models (pgm) [43], which are among the most flexible and principled ways available today for inference from heterogeneous and noisy data. Using a rigorous data pre-processing pipeline in conjunction with this powerful framework, we demonstrate that the (TF, Drug) associations predicted by the method are accurate, by showing that knock-down of the TF affects sensitivity to the drug. The new method is also applicable to other studies where one seeks mechanistic factors underlying individual variation in a quantitative phenotype in the presence of genotype and epigenotype information.

CHAPTER 2: GENMI

2.1 INTRODUCTION

This study, published in Nature Pharmacogenomics in 2016 [44] integrates gene expression, genotype and drug response data in lymphoblastoid cell lines with transcription factor (TF)-binding sites from ENCODE (Encyclopedia of Genomic Elements) in a novel methodology that elucidates regulatory contexts associated with cytotoxicity. The method, GENMi (Gene Expression iN the Middle), postulates that single-nucleotide polymorphisms within TF-binding sites putatively modulate its regulatory activity, and the resulting variation in gene expression leads to variation in drug response. Analysis of 161 TFs and 24 treatments revealed 334 significantly associated TFtreatment pairs. Investigation of 20 selected pairs yielded literature support for 13 of these associations, often from studies where perturbation of the TF expression changes drug response. Experimental validation of significant GENMi associations in taxanes and anthracyclines across two triple-negative breast cancer cell lines corroborates our findings. The method is shown to be more sensitive than an alternative, genome-wide association study-based approach that does not use gene expression. These results demonstrate the utility of GENMi in identifying TFs that influence drug response and provide a number of candidates for further testing.

2.2 METHODS

2.2.1 DATA COLLECTION

We obtained genotype, gene expression and drug response data on 95 Han-Chinese, 96 Caucasian and 93 African American lymphoblastoid cell lines from the Coriell Cell Repository (Camden, NJ, USA). Of these 284 individuals, 176 were females and 108 males, with an average age of 33.44 years. The genotype data consisted of 1 344 658 germline SNPs.

Quality control analysis had already been performed on these SNPs, removing those that deviated from HardyWeinberg equilibrium, were called 95% of the time or had minor allele frequencies of 5%. Gene expression data consisted of 54 613 Affymetrix U133 Plus 2.0 GeneChIP (Santa Clara, CA, USA) probes assayed for the 284 individuals, with raw expression data being transformed using QUOTE GC robust multiarray averaging. Genotype and gene expression data are available at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo`) under SuperSeries acces-

Table 2.1: Treatments analyzed in this study grouped into classes based on pharmacological similarity

Treatment	Family
Paclitaxel	Antimicrotubule Agents (Taxanes)
Docetaxel	
Doxorubicin	Topoisomerase II Inhibitors (Anthracyclines)
Epirubicin	
Cytarabine (ara-C)	Nucleoside Analogues
Gemcitabine	
Triceribine (TCN)	
6-Mercaptopurine (6MP)	Purine Antimetabolites
6-Thioguanine (6TG)	
Cladribine	
Fludarabine	
Carboplatin	Platinum Cores
Cisplatin (CDDP)	
Oxaliplatin	
Everolimus	Mammalian Target of Rapamycin (mTOR) Inhibitors
Rapamycin (Sirolimus)	
Arsenic	Singletons (no drug pair)
Hypoxia	
Metformin	
Medroxyprogesterone Acetate (MPA)	
Methotrexate (MTX)	
N-acetyl-p-benzoquinoneimine (NAPQI)	
Radiation	
Temozolomide (TMZ)	

sion no. GSE24277. These data were published in a study by Niu et al [24].

Drug response data were derived from dosage-response curves of 24 cytotoxic treatments shown in Table 2.1. The phenotype, called EC50, represents the concentration at which the drug reduces the population of LCL cells to half of the initial population. Data for 15 of the 24 treatments have been published in analysis in various studies conducted on these cell lines; in particular, MPA, NAPQI, 6MP, 6TG, araC, oxaliplatin, carboplatin, cisplatin, docetaxel, everolimus, gemcitabine, paclitaxel, metformin, radiation and rapamycin have been analyzed in published studies [24, 45, 46, 47, 48, 49, 50, 51, 42]. Response data for the following nine drugs have not been published: arsenic, cladribine, doxorubicin, epirubicin, fludarabine, hypoxia, MTX, TCN and TMZ. Cytotoxicity assay was performed for every one of these drugs using the LCL panel. After initial optimization, cells were treated with a range of concentrations for any given drug tested, followed by incubation for 48 to 72h. MTS cytotoxicity assays were then performed using Cell Titer 96 AQueous Non-Radioactive Cell Proliferation Assay kit (Promega Corporation, Madison, WI, USA), followed by ab-

sorbance measurement at 490 nm in a Safire2 microplate reader (Tecan AG, Switzerland). Cytotoxicity phenotypes were determined by the best fitting curve using the R package drc (doseresponse curve) (`https://cran.r-project.org/web/packages/drc/drc.pdf`) based on a logistic model.

Experimental data on TF binding were retrieved from the ENCODE project [1]; specifically, the clustered ChIP (version 3) tracks across 91 were used. ChIP tracks consisted of the clustered ChIP peaks of 161 TFs. TF ChIP high occupancy target regions were removed too.

Gene mappings to the Affymetrix arrays were obtained for the Affymetrix Human Genome U133 Plus 2.0 array. ENSEMBL gene symbols were used as the gene reference of choice: we used 55 038 ENSEMBL gene symbols that were annotated with at least one ENSEMBL exon. Of the 54 613 probes assayed on the HG U133 Plus 2.0 array, 37 677 mapped to at least one of the 55 038 ENSEMBL gene symbols.

Human triple-negative breast cancer cell lines, BT549 and MDA-MB231, were obtained from the American Type Culture Collection (Manassas, VA, USA). BT549 cells were cultured in RPMI-1640 containing 10% fetal bovine serum. MDA-MB-231 cells were cultured in L-15 medium containing 10% fetal bovine serum.

2.2.2 DATA PROCESSING

REMOVING HIGH OCCUPANCY TARGET (HOT) REGIONS

We used TF ChIP peaks to focus on SNPs whose association with expression might be mediated by the regulatory action of that TF. However, it is well known that different TFs tend to co-localize at the same genomic loci, a phenomenon that is especially pronounced at HOT (high occupancy target) regions, and a TFs binding at these HOT regions is not necessarily indicative of regulatory function [52]. To enrich the ChIP-based collection of binding sites for functional TF-DNA interactions, we removed segments of 50bp where six or more TFs bind, resulting in a $\sim 25\%$ reduction in the total number of ChIP TF peaks across all ENCODE cell lines.

IMPUTATION PROCEDURE

Imputation was run separately for each race and chromosome separately. Chromosomes were divided into 40MB regions. BEAGLE v3.3.1 was run on these 40MB regions of the genome, plus a 1MB buffer region to the right and left of the main region, as the ends

are generally imputed poorly. The reference and observed genotype data were inputted as phased and unphased, respectively. The lowmem option was used to reduce the overall amount of memory required to run BEAGLE; additionally, the excludemarkers option was used to remove the rare SNPs from the reference mentioned above. For edification, BEAGLE was only used to impute untyped markers, not missing genotyped markers.

PROCEDURE FOR COMPARING GENMI TO THE BASELINE METHOD

In order to compare the number of associations between GENMi and the baseline method, we sought to randomly permute the data set presented to these methods so that we can estimate their false positive rate (FPR). One way to create randomized data sets is to scramble the gene expression to phenotype relationships. However, this kind of randomization is meaningless for estimating FPRs of the baseline method as it does not incorporate gene expression. We required a randomization procedure that allows us to present the same data sets to both methods, thus using the same null distribution for estimating rate of Type 1 errors. Both methods utilized the following three types of information: TF binding sites, genotype, and phenotype, and randomizing one of these three information types would be suitable for defining a null distribution of data sets. We reasoned that permuting the binding locations of TFs would be a more appropriate permutation, as both methods attempt to make the statement that co-location of TF binding locations with impactful SNPs is not-random; randomizing where these peaks are located would destroy the TF-specific information while keeping everything else unchanged and thus test whether the (TF, treatment) associations discovered are indeed specific to the TF.

We randomized the locations of TF peaks by randomizing (SNP, TF peak) assignments, while preserving the relative number of SNPs within each TFs DNA-binding regions, so that it is the identity rather than the total number of SNPs co-locating with TF peaks that is disrupted We segmented the set of all SNPs, X, into two disjoint sets. The first, X_{cis} (called cis-SNPs), is the set of all SNPs within 50kb upstream of any Ensembl gene and represents the set of all cis-SNPs. The second, X_{trans} (called trans-SNPs), is the set of all SNPs not located within any Ensembl gene: $X_{trans} = X - X_{cis}$. For a given TF, we measured the number of cis-SNPs, t_{cis}, and trans-SNPs, t_{trans} within ChIP peaks of that TF. We then generated 5000 random permutations by sampling t_{cis} snps from the uniform distribution on X_{cis} and t_{trans} SNPs from the uniform distribution on X_{trans}. Not only does this destroy the cis-regulatory architecture of a given TF, but it preserves the relative number of SNPs in the cis context (which is important for GENMi) and ensures that each permutation selects the same number of SNPs for a given TF.

2.2.3 CONTROLLING FPR - GENMI TO EXPRESSION-LESS METHODS

For GENMi, the enrichment score (ES) computed by the GSEA procedure on a real data set, for a specific (TF, treatment) pair, was compared to ES values from 5000 randomized data sets for the same TF, thus providing the estimated FPR if this (TF, treatment) pair is called significantly associated.

For methods without expression, we computed the cardinality of intersection between GWAS SNPs with p-value $\leq 10^{-8}$ and the set of SNPs under a TFs peaks. This cardinality, as observed on a real data set, for a given (TF, treatment) pair, was compared to the same statistic observed in 5000 randomized data sets for the same TF, thus providing an estimated FPR. Since this is done within each TF, this is equivalent to an empirical approximation to the hypergeometric test p-value.

2.2.4 EXPERIMENTAL METHODS

All wet-lab experiments were performed in the laboratory of Dr. Liewei Wang at Mayo Clinic, Rochester, MN. RNA interference and qRT-PCR. The small interfering RNAs (siRNAs) for the candidate transcript factors and negative control siRNA were purchased from Dharmacon (Lafayette, CO, USA). Reverse transfection was performed in 96-well plates. Specifically, $3000 - 4000$ cells were mixed with 0.1 ml of lipofectamine RNAi-MAX reagent (Invitrogen, Grand Island, NY, USA) and 10 nM siRNA for each experiment.

Total RNA was isolated from cultured cells transfected with control or specific siRNAs with the Qiagen RNeasy kit (QIAGEN, Valencia, CA, USA), followed by real-time quantitative reverse transcription-PCR (qRT-PCR) performed with the one-step, Brilliant SYBR Green qRT-PCR master mix kit (Stratagene, Santa Clara, CA, USA). Specifically, primers purchased from QIAGEN were used to perform qRT-PCR using the Stratagene Mx3005P Real-Time PCR detection system (Stratagene). All experiments were performed in triplicate with β-actin as an internal control. Reversetranscribed Universal Human reference RNA (Stratagene) was used to generate a standard curve. Control reactions lacked RNA template.

MTS CYTOTOXICITY ASSAY

Epirubicin, doxorubicin, paclitaxel and docetaxel were purchased from Sigma-Aldrich (Milwaukee, WI, USA). Drugs were dissolved in dimethyl sulfoxide and aliquots of stock solutions were frozen at -80°C. Cell proliferation assays were performed in triplicate at each drug

concentration. Cytotoxicity assays with the lymphoblastoid and tumor cell lines were performed in triplicate at each dose. Specifically, 90μl of cells (5×103 cells per ml) were plated into 96-well plates (Corning, Corning, NY, USA) [53] and were treated with 10μl of epirubicin or doxorubicin at final concentrations of $0, 0.0156, 0.03125, 0.0625, 0.125, 0.25, 0.55, 1$ and 2 μmol l-1. Similarly, cells were treated with paclitaxel or docetaxel at $0, 0.01, 0.1, 1, 10, 50, 100, 1000$ and 5000 $n$mol l-1. After incubation for 72 h, 20 μl of CellTiter 96 AQueous Non-Radioactive Cell Proliferation Assay solution (Promega Corporation) was added to each well. Plates were read in a Safire2 plate reader (Tecan AG).

2.2.5 STATISTICAL ANALYSES

SUMMARIZING CYTOTOXICITY CURVES

Significance of the half-maximal inhibitory concentration (IC50) values between negative control siRNA and gene-specific siRNA was determined by the two-tailed unpaired t-test.

STATISTICAL TESTING TF ROLE IN DRUG RESPONSE (GENMI)

We operationally defined the cis-regulatory domain of a gene as the 50-kb sequence upstream of the genes transcription start site. For any given TF, we assigned a TF-specific cis-eQTL score to each ENSEMBL gene as follows:

1. Retain all SNPs located in TF ChIP peaks in the cis-regulatory domain of the gene.
2. Retrieve all HG U133 Plus 2.0 probes mapped to the gene.
3. Compute the correlation (eQTL) for each (SNP, probe) combination. This is the correlation coefficient, across all 284 cell lines, between the SNP genotype and the probes expression value. Also, compute the P-value corresponding to this correlation coefficient.
4. Use the coefficient of determination of the single best eQTL among all (SNP, probe) combinations as the TF-specific cis-eQTL score of the gene. Retain probe and SNP identities contributing to the best eQTL for further analysis.

We then considered the set of 400 genes with the strongest TF-specific cis-eQTL scores (we additionally required that a gene included in this set have a TF-specific cis-eQTL P-value ≤ 0.05, so the cardinality of the set may be ≤ 400). These may be thought of as genes where

genotypic variation in the TFBS correlates with variation in the genes expression, potentially implicating the TF in their expression variation. Although SNPs outside of TFBS can affect TF regulation of gene regulation [50] (for example, SNPs in cofactor binding sites), we limit analysis to eQTL SNPs enveloped within TFBS, as the functional effect of SNPs distant from the TFBS is not well understood. We therefore refer to this gene set as the eQTL gene set of TF. In addition, we required that in order for us to analyze the role of TF in drug response, there should be at least 15 genes with strong eQTLs within binding sites of that TF, that is, the eQTL gene set of TF should have at least 15 genes, as per recommendations accompanying the Gene Set Enrichment Analysis (GSEA) tool. This resulted in the analysis being restricted to 114 of the 161 TFs for which ChIP data were available.

To determine whether the genotype expression associations identified above are linked to the varying response to a given cytotoxic treatment (drug or radiation), we correlated each probes expression value with the EC50 value of the treatment, and ranked all genes by this correlation coefficient. (We used the best correlation coefficient among multiple probes for each gene.) Using GSEA [54], we tested for statistical association between this ranked list and the eQTL gene set of TF defined above. The GSEA procedure reported a P-value that served as the basis for inferring a role for the TF in individual variations in response to the specific cytotoxic treatment. As 114 TFs were separately tested in this manner, we relied upon the false discovery rate (FDR) values reported by GSEA to correct for multiple hypothesis testing with each treatment. We refer to this method as GENMi.

2.3 RESULTS

2.3.1 INTEGRATING GENO/PHENO-TYPE -EXPRESSION & ENCODE DATA

The relationship between genotype and response to cytotoxic treatments is expected to be mediated, at least in part, by regulation of gene expression [17] (Figure 2.1A). Inclusion of expression data in the LCL data set allows us to investigate this hypothesis by simultaneously examining the correlation between genotype and expression and that between expression and phenotype (Figure 2.1B). We hypothesized that SNPs manifesting the genotypeexpression correlation (eQTLs) should reside within binding sites of TFs that orchestrate the transcriptional programs activated or deactivated by the treatment, and that these SNPs influence phenotypic variation through their effect on gene expression [55].

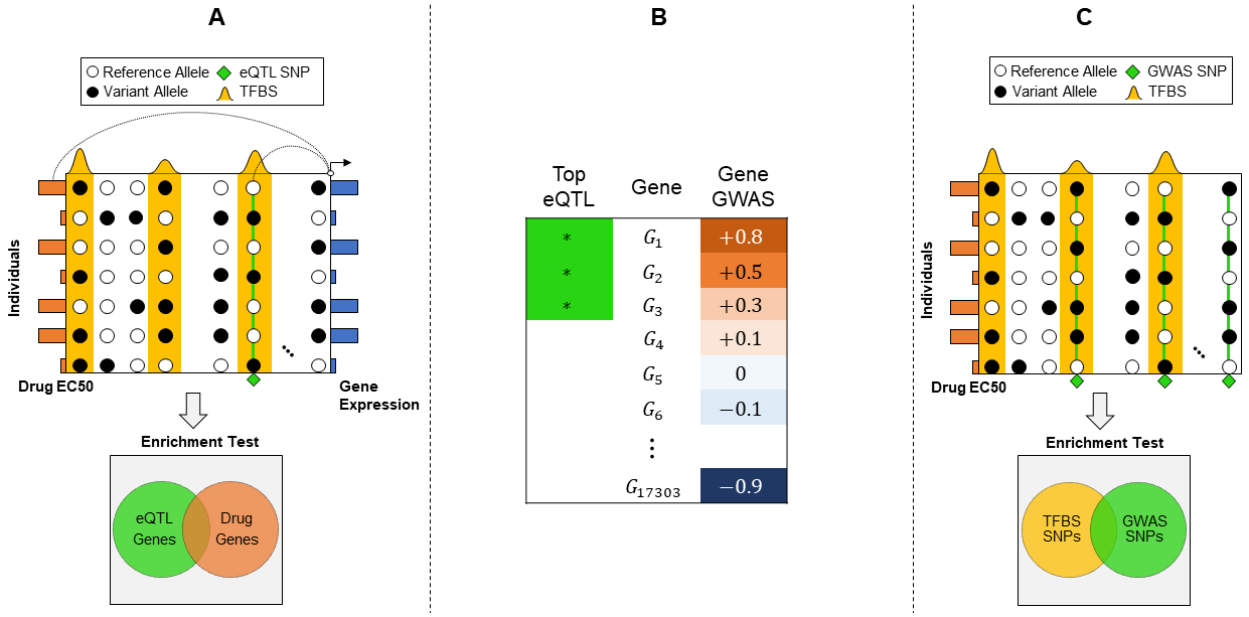


Figure 2.1: (a) The GENMi (Gene Expression iN the Middle) method. Shown is the 50 kb upstream region of a single gene, with transcription factor-binding site (TFBS; ChIP peaks) in yellow, single-nucleotide polymorphisms (SNPs; circles) and their allelic state (black or white) in a sample of seven individuals, as well as gene expression (blue bars on right) and drug response EC50 values (orange bars on left) in these individuals. The gene is scored in two ways: correlation of the best expression quantitative trait locus (eQTL) SNP (green diamond) coincident with a TFBS and correlation of the genes expression with drug response (these two correlations are illustrated by lines connecting the two correlated variables). Integrating over all genes, testing the overlap between strongest eQTL genes and genes associated with drug response (enrichment test, bottom) quantifies the extent to which a TF is associated with drug response via cis-regulatory mechanisms. (b) Cartoon of Gene Set Enrichment Analysis (GSEA) used as the enrichment test in GENMi. Genes are ranked according to their correlation with drug response (gene GWAS). The analysis looks at the extent to which a given gene set (in this case genes carrying the strongest eQTLs coincident with the TFBS) are enriched near the top or bottom of the ranked list. Here, the gene set is strongly associated with genes positively associated with drug EC50 values. (c) Baseline method that does not use expression data. Shown are SNPs (columns) distributed throughout the genome within TFBS (yellow peaks) and outside. Genome-wide association study (GWAS) SNPs (green diamonds) correlated with drug response across individuals (rows) are tested for enrichment with within-TFBS SNPs to determine whether a TF is associated with drug response.

Our goal was to test the possibility that a TF mediates the individual-to-individual variation of gene expression that in turn leads to variation in cytotoxicity across cell lines. To this end, we sought examples (Figure 2.1A) where a SNP inside the TFBS (ChIP peak)

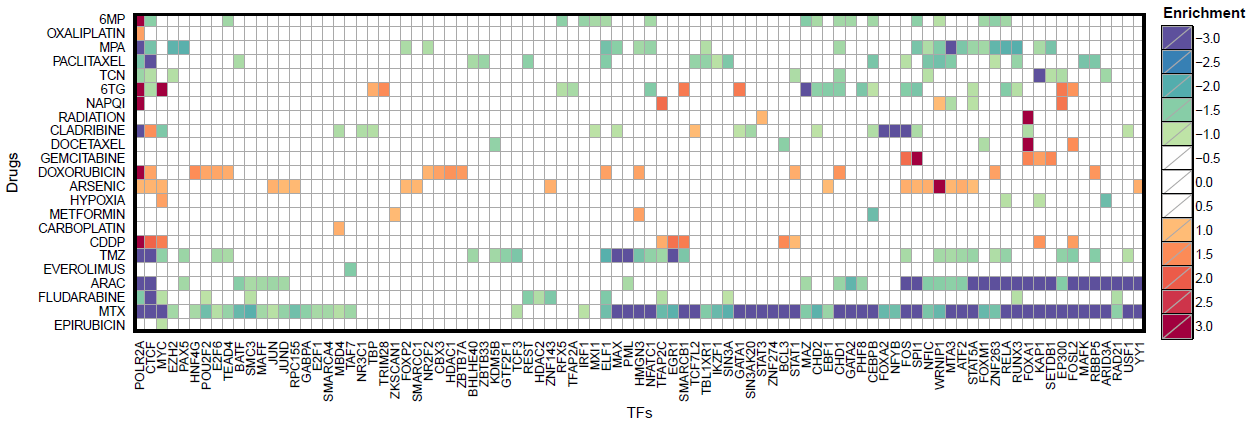


Figure 2.2: Significant (transcription factor (TF), treatment) associations. Shown are the log-transformed false discovery rate (FDR) values for all associations meeting FDR ≤ 0.1. The greenblue range refers to enrichment for genes whose expression negatively correlates with cytotoxicity, and the yellowred range indicates enrichment of genes whose expression positively correlates with cytotoxicity. The yellowred log-transformed FDR values are multiplied by negative 1, creating the 3 to 3 range in the legend enrichment. Anything with an FDR ≥ 0.1 is shown as white.

correlates with the neighboring genes expression (cis-eQTL [56]), and that genes expression correlates with drug response. To formalize this idea as a statistical test (Figure 2.1B), we (1) first ranked genes by the correlation between their expression and the phenotype, (2) separately identified a fixed number of genes (400 in tests reported here) that bear the strongest response. cis-eQTLs within the TFBS and (3) finally used GSEA to test whether the latter set of genes (step 2) is enriched near the top of the former ranked list (step 1). In other words, we asked: when using the TFBS as the context, is genotype-to-expression correlation reflected in expression-to-phenotype correlation? Note that step (1) is performed independently of the TF, and does no hypothesis testing; it simply ranks genes by their (expression) correlation with phenotype. Steps (2) and (3) test whether the cis-eQTLs induced by a TF appear significantly frequently near the top of this phenotype-associated gene list, thus suggesting a role for that TF in the association between genotypic and phenotypic variation, with expression variation in the middle. We call this entire procedure GENMi.

2.3.2 TFS WITH POTENTIAL ROLE IN CYTOTOXIC VARIATION

We used the GENMi method to assign statistical significance, that is, P-value and FDR to each (TF, treatment) pair. In total, drug induced cytotoxicity for 24 drugs were analyzed, of which 9 were prepared specifically for this study.

Table 2.2: Number of TFs associated with each treatment at an FDR of 0.1 using the GENMi method.

Treatment	**# of associations**
Paclitaxel	20
Docetaxel	5
Doxorubicin	16
Epirubicin	1
Cytarabine (ara-C)	35
Gemcitabine	5
Triceribine (TCN)	10
6-Mercaptopurine (6MP)	18
6-Thioguanine (6TG)	23
Cladribine	19
Fludarabine	13
Carboplatin	1
Cisplatin (CDDP)	10
Oxaliplatin	1
Everolimus	1
Rapamycin (Sirolimus)	0
Arsenic	18
Hypoxia	4
Metformin	3
Medroxyprogesterone Acetate (MPA)	24
Methotrexate (MTX)	70
N-acetyl-p-benzoquinone imine (NAPQI)	6
Radiation	2
Temozolomide (TMZ)	29

A total of 3864 pairs were tested (114 TFs $\times$24 treatments).

There are 334 associations at a threshold of FDR ≤ 0.10, involving 91 TFs and 23 treatments (Table 2.2). Figure 2.2 shows all log2 transformed FDR values of any TF and drug with a significant association. The 334 significant associations were distributed unevenly across the treatments, with the drug Methotrexate (MTX) appearing in 70 of the 334 associations (21%), followed by Cytarabine (ara-C) and Medroxyprogesterone Acetate (MPA) as the drugs with most TF associations (see Table 2.2).

Table 2.3: Number of treatments associated for each of 91 out of 114 TFs at an FDR of 0.1 using the GENMi method.

TF	#	**TF**	#	**TF**	#	**TF**	#
POLR2A	15	TFAP2C	5	RAD21	3	ZNF143	2
CTCF	13	ARID3A	4	SIN3A	3	CBX3	1
MYC	9	CHD2	4	SMC3	3	E2F1	1
FOS	8	MAX	4	TBL1XR1	3	GABPA	1
SPI1	8	MAZ	4	TCF7L2	3	GTF2F1	1
WRNIP1	8	SMARCB1	4	YY1	3	HDAC1	1
CHD1	7	STAT1	4	BHLHE40	2	HDAC2	1
ELF1	7	TEAD4	4	FOXA2	2	NR3C1	1
KAP1	7	USF1	4	FOXP2	2	SMARCA4	1
MTA3	7	BATF	3	HNF4G	2	SMARCC1	1
RELA	7	BCL3	3	IKZF1	2	TFAP2A	1
STAT5A	7	E2F6	3	IRF1	2	TRIM28	1
ZNF263	7	EBF1	3	KDM5B	2	ZBTB33	1
CEBPB	6	EGR1	3	MAFF	2	ZBTB7A	1
EP300	6	EZH2	3	MXI1	2	ZKSCAN1	1
FOSL2	6	GATA1	3	NFYB	2	ZNF274	1
FOXA1	6	GATA2	3	NR2F2	2		
NFIC	6	JUN	3	REST	2		
RUNX3	6	JUND	3	RFX5	2		
ATF2	5	MAFK	3	RPC155	2		
FOXM1	5	MBD4	3	SIN3AK20	2		
HMGN3	5	PAX5	3	STAT3	2		
NFATC1	5	PHF8	3	TAF7	2		
RBBP5	5	PML	3	TBP	2		
SETDB1	5	POU2F2	3	TCF3	2		

The TFs with the most numbers of associations at FDR 0.1 (shown in Table 2.3) were POLR2A, the largest subunit of RNA Polymerase II, and CTCF, a versatile regulator involved in gene activation, repression, silencing and chromatin insulation [57]. We have reason to believe (Figure 2.4) that these frequent associations involving general TFs that bind the genome extensively are artifacts of our procedure, in conjunction with linkage disequilibrium and the promiscuous DNA binding of these TFs. We ignored such associations in our follow-up investigations. Also included in the six TFs with the most drug associations were MYC, which plays an important role in reversing multidrug resistance [58], and FOS, a member of the AP-1 complex that is linked to chemotreatment resistance [59].

We next examined the collection of statistically significant (TF, treatment) associations for prior experimental evidence supporting them. To our knowledge, there is no standard

Table 2.4: Literature support for 20 significant (TF, treatment) associations at FDR (false discovery rate) ≤ 0.1 where the TF (transcription factor) is associated with ≤ 5 treatments and the treatment is associated with ≤ 10 TFs.

Association No.	**Treatment**	**TF**	**Literature Evidence**
1	Cisplatin	EGR1	Direct
2	Cisplatin	STAT1	Direct
3	Docetaxel	FOXM1	Direct
4	Radiation	STAT3	Direct
5	Cisplatin	SMARCB1	Direct
6	Cisplatin	BCL3	Direct
7	TCN	EZH2	Indirect
8	Gemcitabine	SETDB1	Indirect
9	Docetaxel	KDM5B	Indirect
10	Hypoxia	ARID3A	Indirect
11	Carboplatin	MBD4	Indirect
12	Cisplatin	TFAP2C	Indirect
13	Docetaxel	BCL3	Indirect
14	Everolimus	TAF7	None
15	NAPQI	TFAP2C	None
16	TCN	ARID3A	None
17	TCN	STAT1	None
18	Metformin	HMGN3	None
19	Metformin	ZKSCAN1	None
20	TCN	SETDB1	None

benchmark that can help us with such an assessment, and hence we resorted to surveying the literature for studies implicating a TF in the response to a specific cytotoxic treatment, for example, TFs whose knockdown or overexpression has been shown to affect cytotoxicity, though not necessarily in the lymphoblastoid cell line. We focused on significant (TF, treatment) associations that are relatively unique, that is, the TF is associated with ≤ 5 (of 24) treatments and the treatment is associated with ≤ 10 (of 114) TFs. These 20 associations are shown in Table 2.4. We noted 6 of the 20 associations to be supported by direct experimental evidence involving the drug and the TF. We discuss these below.

2.3.3 LITERATURE ASSOCIATIONS

DIRECT ASSOCIATIONS

Observation: FoxM1 (transcription factor forkhead box protein M1) is associated with response to docetaxel. Remarks: overexpression of FoxM1 in gastric cancers was previ-

ously shown to mediate resistance to docetaxel and inhibiting FoxM1 was found to reverse docetaxel resistance in gastric cancers [60]. Similar conclusions were reached by other studies [61].

Observation:EGR-1 (early growth response protein 1) is associated with cisplatin treatment. Remarks: EGR-1 has been shown to regulate cisplatin-induced apoptosis in human esophageal squamous cell carcinoma cell lines (WHCO1) [62]. The EGR-1 promoter has been shown to be induced by this drug [63, 64].

Observation: STAT1, a member of the signal transducer and activator family of transcription factors, is associated with cisplatin. Remarks: overexpression of STAT1 in A2780 human ovarian cancer cells was shown to increase cisplatin resistance [65]. Moreover, inhibiting STAT1 expression has been shown to attenuate cisplatin-induced ototoxicity in rats [66] and mice [67].

Observation: STAT3, a homolog of STAT1 in the signal transducer and activator family of transcription factors, is associated with radiation treatment. Remarks: a previous tudy found STAT3 blockade to enhance radiosensitivity in Hep-2 cells [68]. Other studies have reported that radiation activates STAT3 [69], and that targets of STAT3 are upregulated by radiation in a mouse model of glioblastoma [70].

Observation: SMARCB1, a core component of the switch/sucrose nonfermentable (SWI/SNF) nucleosome remodeling complex, is associated with cisplatin. Remarks: recent sequencing of various cancer cells have demonstrated frequent mutations in SWI/SNF factors such as ARID1A. Suppression of ARID1A and its paralog ARID1B sensitized the cell to cisplatin as well as radiation. Suppression of SMARCB1 reproduced the same effects [71].

Observation: BCL-3 is associated with cisplatin. Remarks: a previous study found BCL-3 overexpression to suppress cisplatin-induced apoptosis in MCF7 breast cancer cell lines [72].

In addition to the above six examples of experimental results directly supporting an association, we also noted seven of the statistical associations from Table 2.4 to be supported by indirect experimental evidence involving transcriptional regulation of the TF in response to the drug or direct experimental evidence involving a protein closely related to the TF, shown below:

INDIRECT ASSOCIATIONS

The following seven (TF, treatment) associations were noted as having indirect support from the literature, i.e., evidence that falls short of a direct manipulation of the TFs expression resulting in changed drug response.

Observation: KDM5B, a histone lysine demethylase, is associated with docetaxel. Re-

marks: Interestingly, the KDM5B homolog KDM5A has been noted to have higher expression in breast cancer patients with improved response to neoadjuvant docetaxel treatment [73]. The two closely related homologs KDM5B and KDM5A have been shown to highly conserved mechanism of histone recognition and binding [74], and it is thus possible that the statistical association discovered using ChIP data on KDM5B reflects the regulatory role of KDM5A in docetaxel response.

Observation: The histone methyltransferase SETDB1 is associated with gemcitabine. Remarks: A recent study showed SETDB1 to be down-regulated by the action of p53 during cell death induced by drug paclitaxel [75]. A different study showed that gemcitabine treatment induces accumulation of p53 in cell lines [76], the two pieces of evidence together providing indirect support for the noted statistical association.

Observation: ARID3A, a B-cell regulator of immunoglobulin heavy-chain transcription (BRIGHT), is associated with hypoxia. Remarks: The micro RNA MiR-125b has been shown to be significantly upregulated in endothelial cells under hypoxic conditions [77]. A separate study using both murine and human B-cells shows that MiR-125b is a direct down regulator of ARID3a and a plausible mechanism for resistance to apoptosis [78].

Observation: EZH2, a histone lysine methyltransferase from the mTOR/AKT pathway, is associated triciribine. Remarks: Suppression of EZH2 has been associated with increased sensitivity to chemotreatment in human hepatocellular carcinoma (HCC) cells [79].

Observation: MBD4, a DNA repair and methyl CpG binding protein, is associated with carboplatin. Remarks: Previous studies found silencing of MBD4 to increased sensitivity to oxaliplatin, a platinum treatment similar to carboplatin, in the HCT 116 and SW480 tumor cell lines [80]. Overexpression of MBD4 was found to alter colony survival upon treatment with cisplatin, the parent compound of carboplatin, in a human colon carcinoma cell line [81].

Observation: TFAP2C, a member of the Activating Protein 2 (AP-2) transcriptional family, is associated with cisplatin. Remarks: siRNA knockdown of the homologous protein TFAP2A was previously shown to decrease cisplatin sensitivity in bladder cancer [82]. TFAP2-A and TFAP2-C share 76% similarity in their C terminal halves that harbor the DNA binding domain and are known to bind to the GTGACGTCAG consensus motif 9113991 [83].

Observation: BCL-3, a proto-oncogenic transcription factor in B-cell lymphoma, is associated with docetaxel. Remarks: Docetaxel sensitivity has been found to increase upon inhibition of NF-κB [84], which is a direct driver of BCL-3 expression [87]. Additionally, overexpression of BCL-2, another B cell lymphoma proto-oncogene, was linked to docetaxel sensitivity in small non-small cell cancer [85].

Table 2.5: Associations from Table 2.4 with direct or indirect support, re-evaluated for each sub-population (AA=African American, CA=Caucasian American, HCA=Han Chinese American). P-values are shows for each sub-population, with and without the use of imputed SNPs. P-values <= 0.05 are shown in bold.

		P-val (no imputation)			**P-val (imputation)**		
DRUG	**TF**	**AA**	**CA**	**HCA**	**AA**	**CA**	**HCA**
CDDP	EGR1	0.919	**0.006**	0.607	**0.009**	0.139	**0.018**
CDDP	BCL3	0.654	0.320	0.010	0.891	0.125	**0.005**
CDDP	SMARCB1	0.454	0.088	0.032	0.051	0.079	**0.002**
CDDP	STAT1	0.072	0.645	0.787	0.884	0.675	0.889
CDDP	TFAP2C	0.705	0.782	0.160	0.093	0.266	**0.006**
Docetaxel	BCL3	0.265	0.677	0.514	**0.035**	**0.012**	0.283
Docetaxel	FOXM1	**0.010**	**0.000**	0.396	**0.000**	0.065	0.851
Docetaxel	KDM5B	**0.004**	0.738	0.304	**0.000**	**0.001**	0.135
Carboplatin	MBD4	**0.008**	**0.044**	**0.007**	0.277	0.705	0.089
Gemcitabine	SETDB1	0.092	0.085	0.058	0.590	0.247	0.137
Hypoxia	ARID3A	0.511	**0.015**	**0.018**	0.386	0.300	0.117
Radiation	STAT3	0.450	0.642	0.655	0.076	0.740	0.886

For seven of the associations noted in Table 2.4, we were unable to find strong supporting evidence from the literature, making these promising candidates for future experimental follow-up. To test the effects of imputation on these results, we replicated the GENMi pipeline using both imputed and genotyped SNPs, to see how many of the literature-supported associations in Table 2.4 were corroborated. The results are shown in Table 2.5. Overall, imputation did not significantly alter the associations reported in Table 2.4: we found 10 of the 13 associations with literature support reported in Table 2.4 to be recovered in this new analysis at the nominal P-value threshold of 0.05 (FDR $\leq$ 0.13). In additionally, form of literature support, 9 were significant in at least one sub-population, at a nominal P-value of 0.05, using either the genotyped SNPs or genotyped as well as imputed SNPs.

2.3.4 POPULATION SPECIFIC ANALYSIS

We tested the effect of population stratification by repeating the genotyped SNP GENMi and genotyped with imputed SNP GENMi analyses on each population separately. The populations studied were Han-Chinese American (HCA, n = 95), African American (AA, n = 93), and Caucasian American (CA, n = 96). Since we have three populations for each analysis, we require that at least on population have the association at the requisite

significance level. The nominal p-values for the genotyped SNP GENMi and genotyped with imputed SNP GENMi analyses are located in Table 2.5. The results for each method are discussed in the following sections.

For genotyped SNPs, a nominal p-value of 0.05 recovered 7 out of 13 associations. Specifically, 4 out of 6 and 3 out of 7 of the direct and indirect associations. Three associations were replicated across multiple populations: (Carboplatin, MBD4), (Docetaxel, FOXM1), and (Hypoxia, ARID3A) in (AA, CA, HCA), (AA, CA), and (CA, HCA) respectively. The other 4 associations were specific to populations: 1 association in AA, 1 association in CA, and 2 associations in HCA.

Similar to the analysis using only genotyped SNPs, a nominal p-value of 0.05 recovered 7 out of 13 associations. Though carboplatin and hypoxia associations were lost when including imputed SNPs, more associations for docetaxel and CDDP were gained, namely with BCL3 and TFAP2C respectively. Two associations were replicated across multiple populations: (Docetaxel, KDM5B) and (Docetaxel, BCL3) in (AA,CA) and (AA, CA). The other 5 associations were population specific: 2 AA associations and 4 HCA associations.

2.3.5 SIMILAR METHOD SANS EXPRESSION REPORTS FEWER ASSOCIATIONS

To determine the utility of our method that integrates genotype, gene expression and phenotype to identify (TF, treatment) pairs, we devised a baseline method agnostic of gene expression. This baseline method (Figure 2.1C) tests whether GWAS SNPs (P-value $\leq 10^{-8}$) for a given treatment are enriched within peaks of a particular TF, computing a P-value of association for each (TF, treatment) pair. We sought to compare the number of significant associations discovered by GENMi and the baseline method respectively at a fixed false positive rate.

Table 2.6: Number of (TF, treatment) associations discovered by the GENMi method and the baseline method that does not use expression data, at varying FPR threshold. The FPR is estimated by running either method on 5000 randomized data sets where transcription factor-binding site (TFBS) locations have been shuffled genome wide.

FPR	**1**	$\mathbf{2 \times 10^{-1}}$	$\mathbf{2 \times 10^{-2}}$	$\mathbf{2 \times 10^{-3}}$	$\mathbf{2 \times 10^{-4}}$
Baseline	1932	75	16	2	0
GENMi	2736	943	211	33	14

Table 2.7: Shown are the functional validation results for 21 TFs enriched (at FDR 0.1 and P-value ≤ 0.05) for either taxane, paclitaxel (PAX) or docetaxel (DOC).

#	TF	GENMi Enrichments				Cell Lines			
		PAX		DOC		MDA-MB-231		BTF549	
		P-Value	FDR	P-Value	FDR	PAX	DOC	PAX	DOC
1	BATF	0.0142	0.07	0.5160	1.00	UP	UP		
2	BCL3	0.4112	0.56	0.0020	0.03	UP	UP	UP	UP
3	BHLHE40	0.0088	0.06	0.0487	0.30	UP	UP		
4	CEBPB	0.0024	0.02	0.3199	0.86	UP	UP	UP	UP
5	ELF1	0.0020	0.03	0.0147	0.20	UP	UP	UP	UP
6	FOS	0.0127	0.09	0.2007	0.63	UP	UP		
7	FOSL2	0.0285	0.12	0.0012	0.03	UP	UP		
8	FOXM1	0.2754	0.44	0.0052	0.07	UP	UP	UP	
9	IKZF1	0.0213	0.09	0.1422	0.78	UP	UP		
10	KDM5B	0.0016	0.02	0.0004	0.04	UP	UP		
11	MAFK	0.0016	0.02	0.6006	1.00	UP	UP	UP	UP
12	MTA3	0.0027	0.03	0.1556	0.62	UP	UP	UP	
13	NFIC	0.0020	0.02	0.0131	0.18	UP	UP	UP	UP
14	RBBP5	0.0012	0.02	0.3613	0.69	UP			
15	REST	0.0008	0.04	0.2343	0.57	UP	UP	UP	UP
16	SIN3A	0.0028	0.03	0.5193	0.78	UP			
17	TBL1XR1	0.0071	0.04	0.5347	1.00				
18	TCF7L2	0.0056	0.04	0.1004	0.45	UP			
19	WRNIP1	0.0004	0.02	0.2273	0.59	UP	UP	UP	
20	ZBTB33	0.0044	0.05	0.9282	0.95	UP	UP		
21	ZNF263	0.0168	0.09	0.0255	0.26	UP	UP		

Validation was performed in two triple negative breast cancer cell lines (BTF549 and MDA-MB-231). For each cell line, drug and TF, a small interfering RNA (siRNA) knockdown experiment was performed, followed by an MTA assay for the drug. Comparisons were made to negative siRNA experiments to determine whether the TF decreased, increased or did not affect the sensitivity of the cell to the drug. UP in the table refers to decreased sensitivity or desensitization of the cell to the drug, that is, the TF knockdown increased cell resistance/survivability to increasing concentrations of the apoptotic drug. In no case was the knockdown found to decrease cell resistance. Cells with P-value ≤ 0.05 or FDR ≤ 0.1 are colored gray and represent the drug for which the TF was predicted to influence response.

Table 2.8: Shown are the functional validation results for 14 TFs enriched for either anthracycline drug, epirubicin (EPI) or doxorubicin (DOX) at FDR 0.1 and P-value ≤ 0.05.

#	TF	GENMi Enrichments				Cell Lines			
		DOX		EPI		MDA-MB-231		BTF549	
		P-Value	FDR	P-Value	FDR	DOX	EPI	DOX	EPI
1	CBX3	0.0033	0.05	0.4734	0.77	UP	UP		
2	CHD1	0.0036	0.04	0.5706	0.82				
3	E2F6	0.0051	0.06	0.0944	0.76				
4	ELF1	0.0008	0.05	0.303	0.79	UP	UP	UP	UP
5	HDAC1	0.002	0.04	0.9026	0.89	UP	UP	UP	UP
6	HMGN3	0.0077	0.05	0.4534	0.76	UP	UP		
7	HNF4G	0.0004	0.03	0.5844	0.82	UP	UP	UP	UP
8	MYC	0.1112	0.32	0.0028	0.09	UP	UP	UP	
9	NR2F2	0.0112	0.08	0.8199	0.82	UP	UP		
10	POU2F2	0.0048	0.06	0.1192	0.72	UP	UP		
11	RBBP5	0.0004	0.04	0.7434	0.83				
12	STAT1	0.0065	0.07	0.7122	0.87	UP	UP		
13	TEAD4	0.0101	0.07	0.8462	0.98				
14	ZNF263	0.0071	0.05	0.1045	0.72	UP	UP	UP	UP

Validation was performed in two triple negative breast cancer cell lines (BTF549 and MDA-MB-231). For each cell line, drug and TF, a small interfering RNA (siRNA) knockdown experiment was performed, followed by an MTA assay for the drug. Comparisons were made to negative siRNA experiments to determine whether the TF decreased, increased or did not affect the sensitivity of the cell to the drug. UP in the table refers to decreased sensitivity or desensitization of the cell to the drug, that is, the TF knockdown increased cell resistance/survivability to increasing concentrations of the apoptotic drug. In no case was the knockdown found to decrease cell resistance. Cells with P-value ≤ 0.05 or FDR ≤ 0.1 are colored gray and represent the drug for which the TF was predicted to influence response.

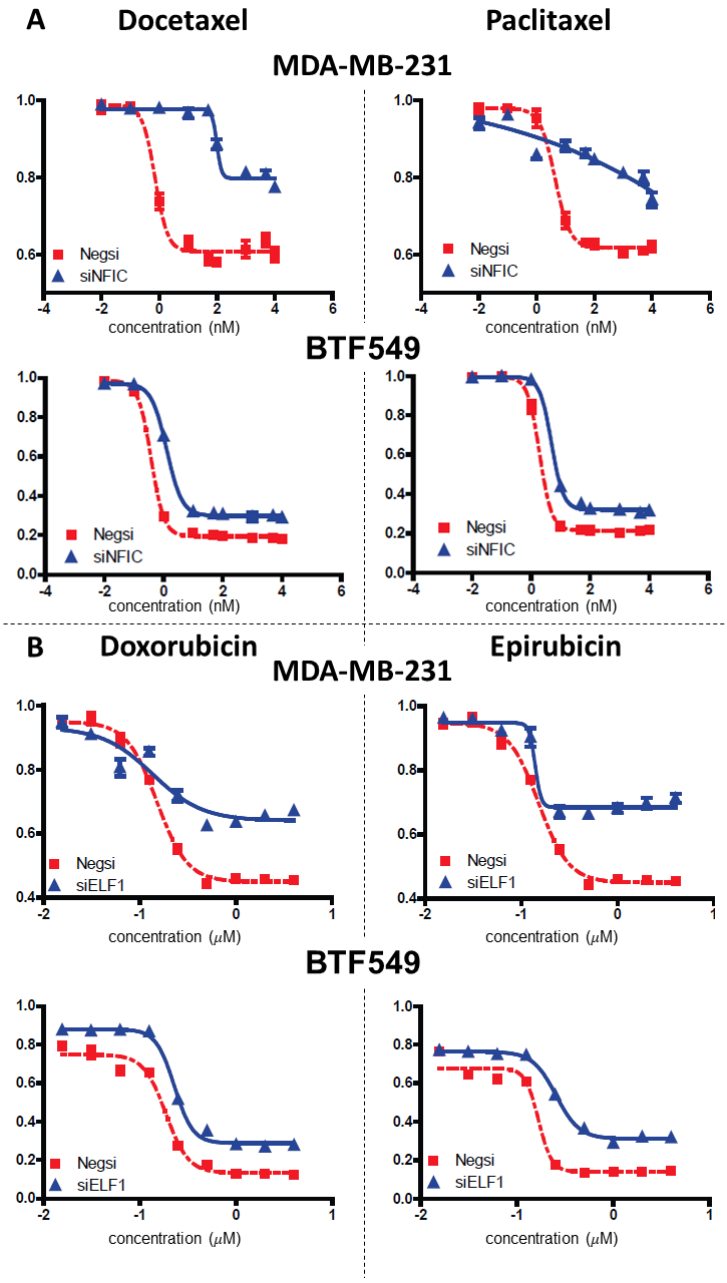


Figure 2.3: (a) Dosage-response curves for the transcription factor (TF) NFIC across two drugs, docetaxel (left) and paclitaxel (right), and two cell lines, MDA-MB-231 and BTF549. Each plot shows significant increase in resistance to the drug upon knockdown of NFIC compared with normal response of the cells, using a two-tailed paired t-test. (b) Dosage-response curves for the TF ELF1 across two drugs, doxorubicin (left) and epirubicin (right), and two cell lines, MDA-MB-231 and BTF549. Each plot shows significant increase in resistance to the drug upon knockdown of NFIC compared with normal response of the cells, using a two-tailed paired t-test.

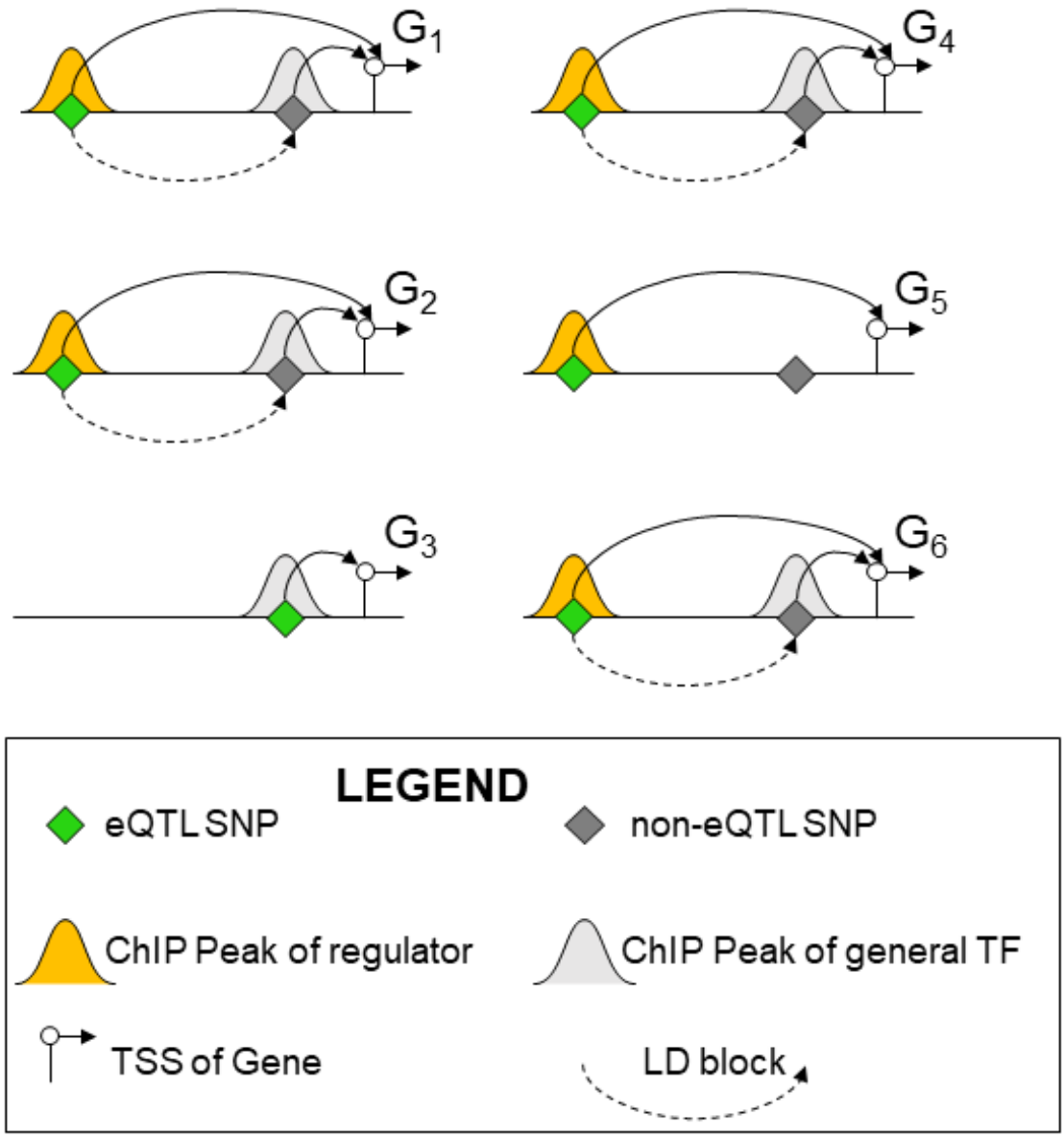
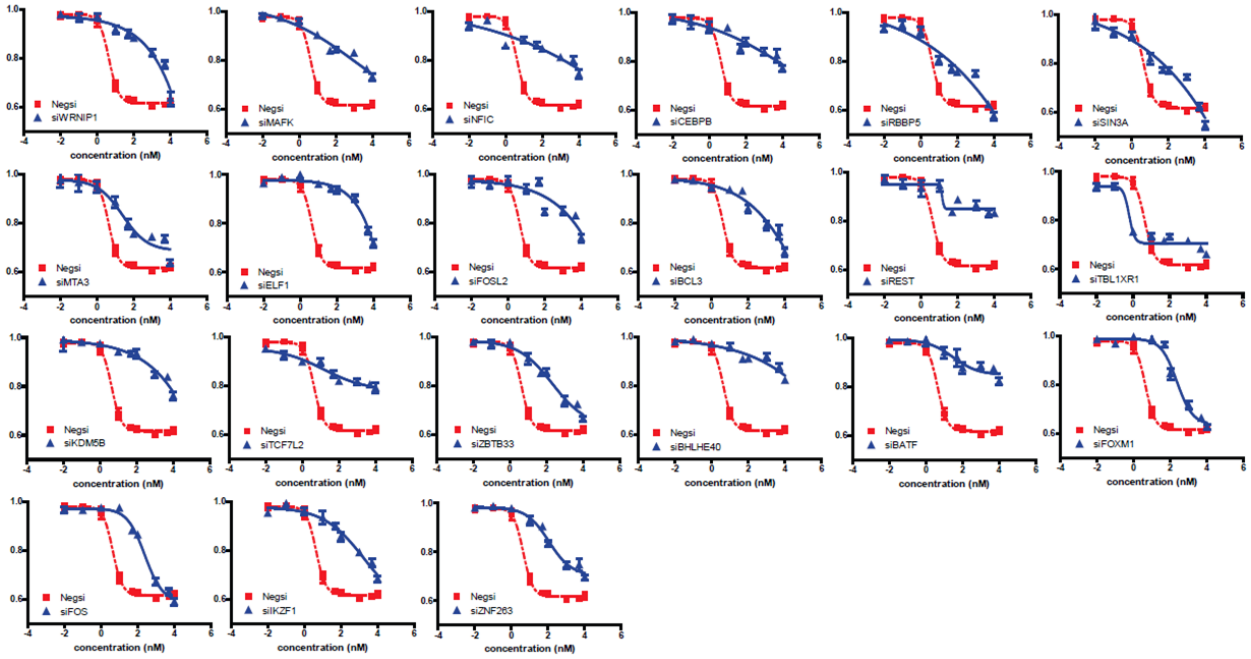
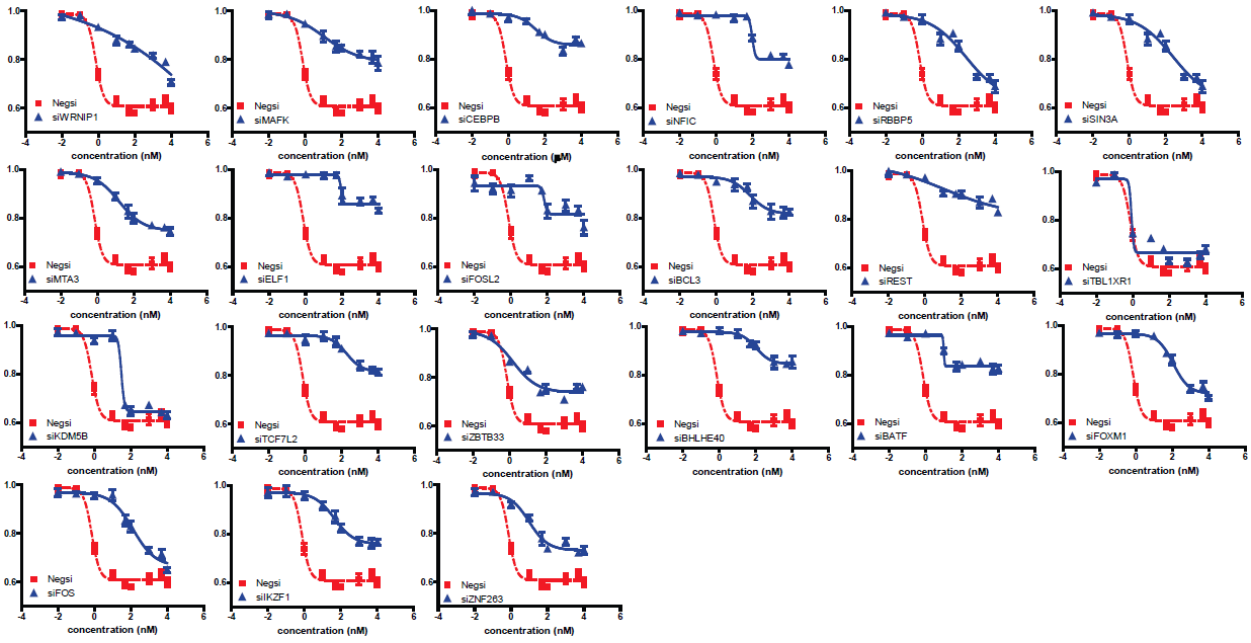


Figure 2.4: Six different gene loci are shown. The true regulator of the phenotype operating through gene expression is illustrated as a yellow peak proximal to the TSS of a gene. A general TF ChIP peak, important for gene expression and present for all transcribed genes, is shown illustrated as a gray peak. An eQTL SNP affecting the regulators activity is shown in green for each locus. Due to a linkage disequilibrium between markers in local proximity (which is likely to be true for a 50kb cis-regulatory region), the non-eQTL SNPs, shown in dark gray, within the general TFs peak is likely to replicate the signal of the eQTL SNP in the true regulators peak due to linkage disequilibrium between markers in the same local region which is a safe assumption for a 50kb cis-regulatory region. In this way, if the regulator is associated with the drug response, the general TF will be as well since it is essentially copying the regulatory signal from the true regulator.

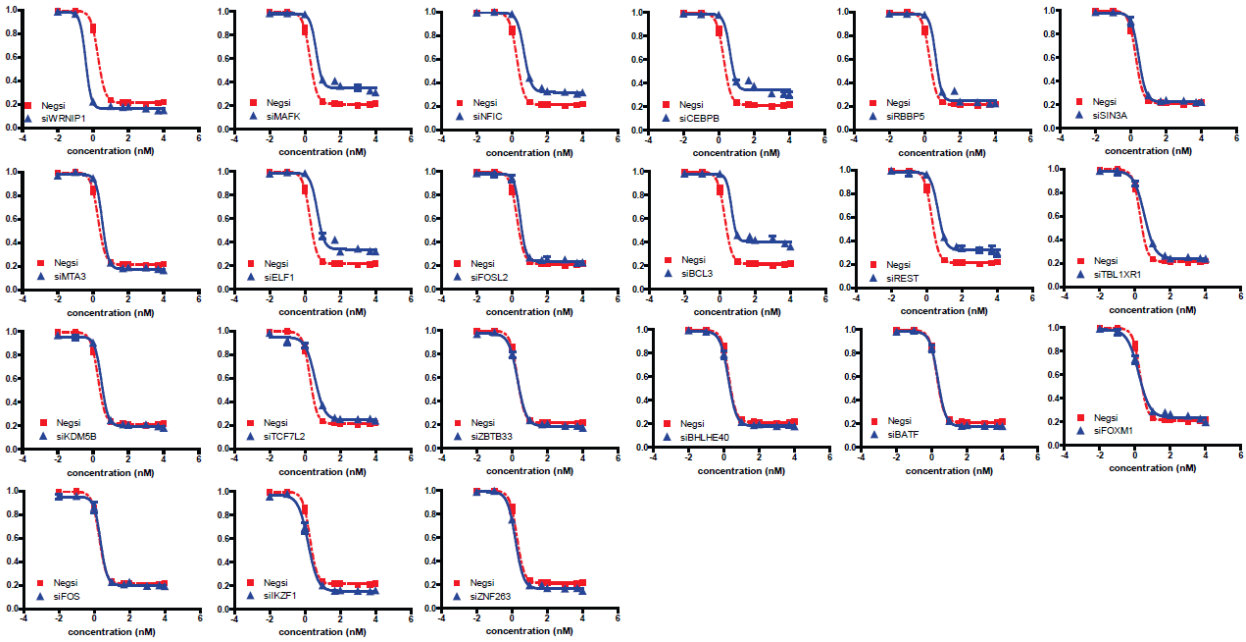


(a) Paclitaxel

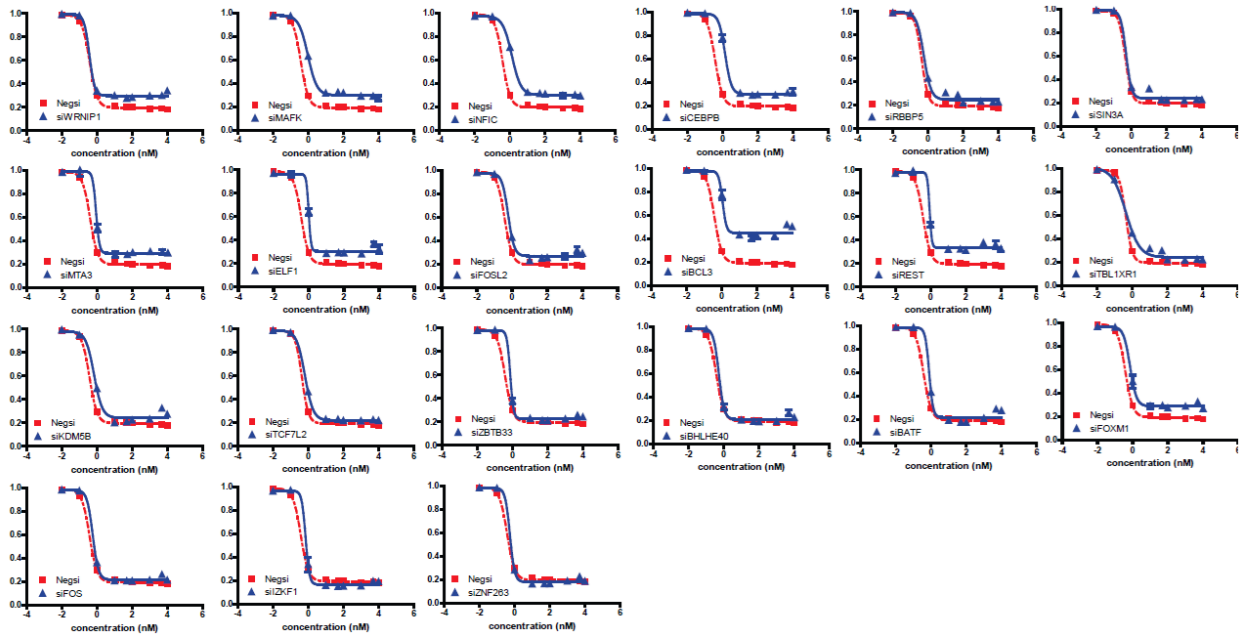


(b) Docetaxel

Figure 2.5: Knockdown cytotoxicity dosage-response curves of 21 TFs with **(A)** paclitaxel and **(B)** docetaxel drugs in the MDA-MB-231 triple negative breast cancer cell line. Concentration is measured in nM.

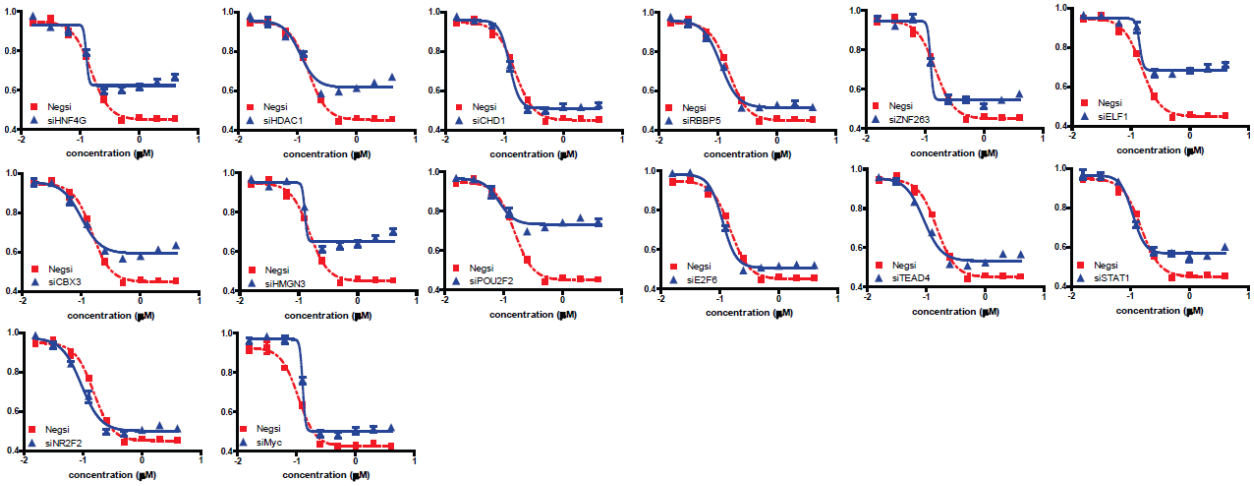


(a) Paclitaxel

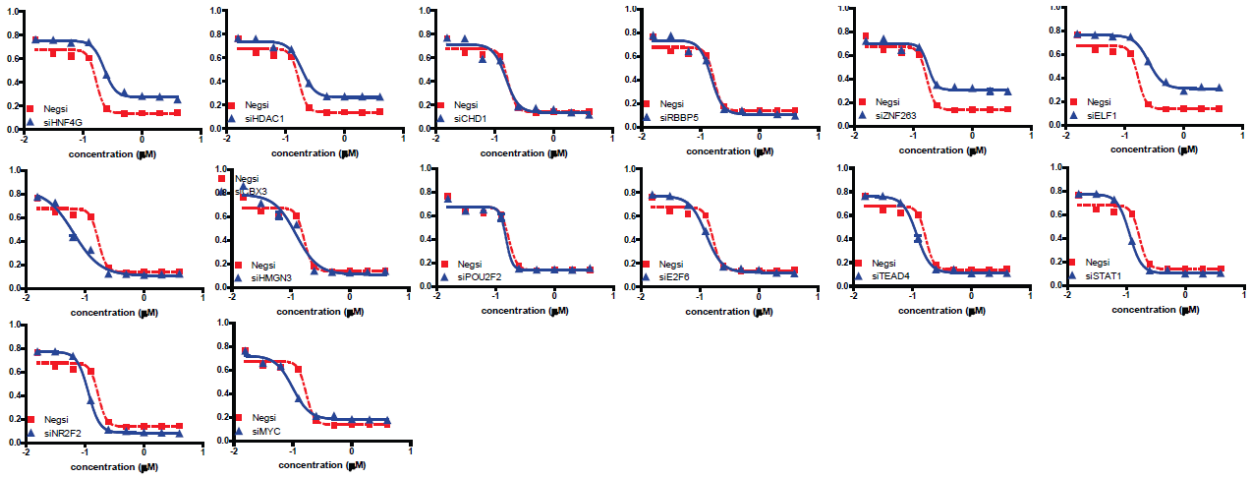


(b) Docetaxel

Figure 2.6: Knockdown cytotoxicity dosage-response curves of 21 TFs with **(A)** paclitaxel and **(B)** docetaxel drugs in the BTF549 triple negative breast cancer cell line. Concentration is measured in nM.

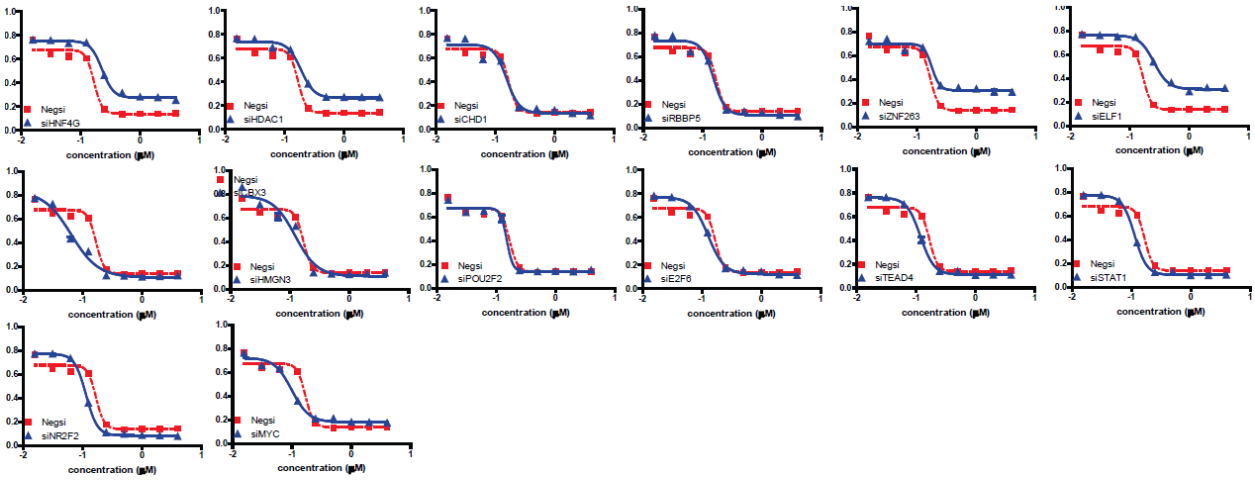


(a) Epirubicin

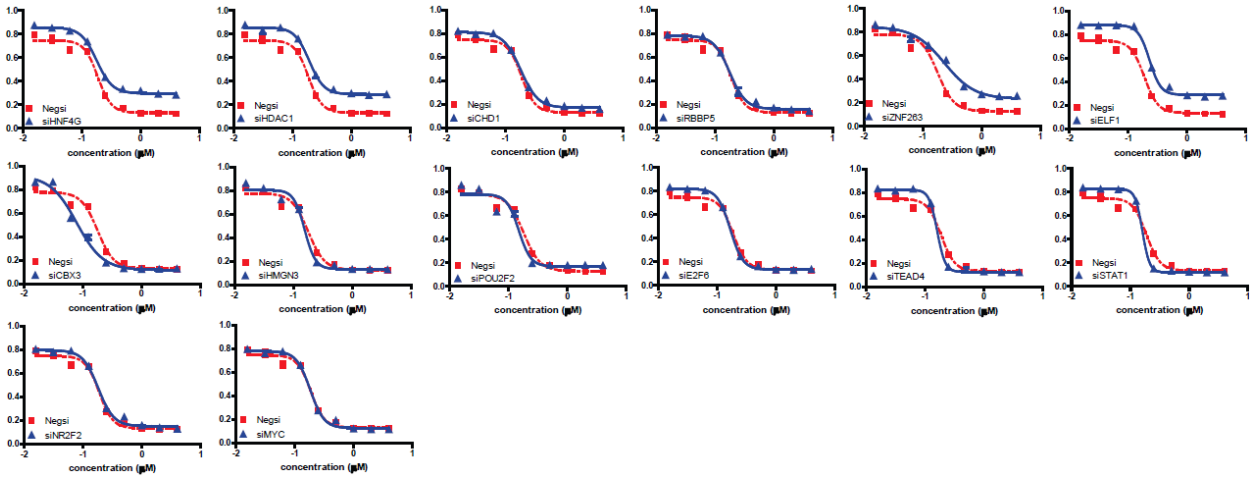


(b) Doxorubicin

Figure 2.7: Knockdown cytotoxicity dosage-response curves of 14 TFs with **(A)** epirubicin and **(B)** doxorubicin drugs in the MDA-MB-231 triple negative breast cancer cell line. Concentration is measured in nM.



(a) Epirubicin



(b) Doxorubicin

Figure 2.8: Knockdown cytotoxicity dosage-response curves of 14 TFs with **(A)** epirubicin and **(B)** doxorubicin drugs in the BTF549 triple negative breast cancer cell line. Concentration is measured in nM.

For a fair comparison, we devised a procedure that generates randomized data sets and asked whether either method discovers a significant association (a false positive, as the data set is a randomized one) on it. By performing this test repeatedly and counting how frequently each method (GENMi or the baseline method) reported false associations, we were able to control for the false positive rate of each method in exactly the same manner. The number of (TF, treatment) pairs reported by either method on the real data set, at each false positive rate threshold, is shown in Table 2.6. The number of significant associations found by the GENMi procedure far outweighs those in the baseline, indicating that utilizing expression information improves the sensitivity of the association study.

2.3.6 IN VIVO EXPERIMENTAL VALIDATION OF TFS

We sought to verify whether TFs associated with drug response variation can be linked in vivo to significant changes in cellular sensitivity to drug-induced apoptosis. Though we utilized lymphoblastoid cell lines data for our association analysis, we performed siRNA knockdown experiments in two different cell lines to demonstrate the generalizability of our results and the LCL model system. Specifically, we choose two triple-negative breast cancer cell lines, BTF549 and MDA-MB-231, that are of great clinical significance. In addition, we restricted our analysis to two of the most widely utilized family of drugs used in the treatment of breast cancer: anthracyclines (doxorubicin, epirubicin) and taxanes (docetaxel, paclitaxel). In addition to being clinically relevant, the mechanisms of the drugs within each family are very similar, gifting us with the ability to check self-consistency within a drug family, that is, if TF increases resistance to doxorubicin, it should also increase resistance to epirubicin. Cost and time constraints restricted the experimental validation in this study to these four drugs.

To choose candidate TFs for validation, we restricted ourselves to those TFs that exhibited a P-value of ≤ 0.05 and FDR of ≤ 0.1 for at least one drug in the family of interest. For the taxanes, this produced 21 TFs, as shown in Tables 2.7 - 2.8 of which were omitted for various reasons. For the anthracyclines, these criteria yielded 14 TFs, as shown in Table 2.4. CTCF and POLR2A were original candidates but omitted because they are ubiquitous activators for expression. The siRNA knockdowns were performed for the 21 taxane- and 14 anthracycline-associated TFs with negative siRNA as a control.

Figures 2.5-2.8 show the results for each cytotoxicity experiment while Tables 2.7-2.8 show the results of the assays for the taxanes and anthracyclines, respectively. Even though a TF was tested even if it was predicted to be associated with only one of the two drugs in a family, our definition of successful validation conservatively required a TF to affect significant change

in the dosage-response curve for both drugs in the family. For additional stringency, this requirement had to be met in both the tested cell lines. Using these stringent criteria, we found 6 out of the 21 predicted TFs, specifically BCL3, CEBPB, ELF1, MAFK, NFIC and REST, to increase resistance to taxane-induced cytotoxicity. An example of what constitutes significant change (induced by a TF knockdown) in the dosage-response curve for the taxanes is shown in Figure 2.3A, for NFIC knockdown. For the anthracyclines, we found 4 out of the 14 predicted TFs, namely ELF1, HDAC1, HNF4G and ZNF263, to increase cytotoxic resistance to doxorubicin and epirubicin in both tested cell lines. An example of a validated TF association (ELF1) for the anthracyclines is shown in Figure 2.3B. The rest of the cytotoxicity curves are shown in Figures 2.5-2.8. In addition, the GENMi analysis predicted MYC to be associated only with epirubicin, and the experimental validation supports this as it is associated only with epirubicin (in BTF549).

Although several of GENMi associations were not corroborated experimentally in both drugs within a family and in both cell lines, this is expected to an extent as the selection of TF knockdowns was based on GENMi predictions for either one of the drugs in the family and made from a different cell line, that is, the experimental test was more stringent than what the statistical association suggests. In total, we find the hit rates of 621 and 414 as significant evidence that GENMi identifies TFs that truly regulate cellular response to drug-induced apoptosis.

2.4 DISCUSSION

We have presented a methodology for interrogating the extent to which specific TFs are associated with individual variations in drug-induced cytotoxicity. We employ a statistical approach that assumes cis-regulatory variants embedded within TFBS affect proximal genes, whose varying expression is then reflected in drug response. Our approach is fundamentally different from the triangulation approach in that we integrate TFDNA-binding data and are able to associate TFs with drug response; also, we do not require the direct association of SNPs with drug-induced phenotype. Focusing on drugs and TFs that feature in a limited number of associations, we noted the statistically significant (TF, treatment) associations to be frequently supported by the literature, in the form of experiments where knockdown or overexpression of the TF changes drug response. Stringent control of the randomization procedure illuminated the benefit of GENMi over a simple GWASTFBS overlap approach where expression data are not used. Although our results showcase the true positive rate of GENMi, they do not yet allow us to determine false negative rates or sensitivity because of the absence of a comprehensive benchmark of true (TF, treatment) associations. Never-

theless, our methods represent the first comprehensive methodology for assessing regulatory associations with drug response.

There are several areas in which the GENMi method could be improved. For one, differences in allele frequencies between stratified populations has been shown to induce spurious associations [86]; adjusting for this confounding factor in a more principled framework may reduce the number of false positives in our results. Second, although filtering for ChIP high occupancy target regions helps eliminate regions where it is difficult to assign function to any one TF, one may not assume that all bound TFs are nonfunctional [52], and future extensions of our method will be cognizant of this. The literature also indicates the existence of eQTL hot spotseQTLs associated with a large number of genes as a result of various confounding factors; elimination of the factors would aid in the discovery of true eQTL signals [87]. Another way of improving GENMi would be to exploit prior knowledge of the relationship between drugs; relevant methods already exist, to an extent [88]. GENMi could also be improved at the level of determining gene targets for a TF. At a statistical level, GENMi reduces to a two-step procedure of enriching top cis-eQTL genes under a given transcriptional regulatory context with genes whose expression correlates with a drug. Ideas from a recent study that employs a two-step regressive framework to a similar end (but without the integration of TF ChIP data)[17] may be adapted to eliminate the arbitrary threshold of the GENMi method in determining transcription gene targets. In addition, GENMi could be improved by considering more elegant eQTL models, such as the methods employed by Sudarsanam and Cohen [89]; however, exploiting more complex methods capturing multi-additive and epistatic interactions at the genome-wide level carries a heavy computational price that is not easy to circumnavigate. Another area of investigation involves the determination of the cis-regulatory region: although cis-regulatory eQTLs are replicated better across studies than trans-eQTLs [90], definitions of cis-regulatory regions differ widely [94]. In our analysis, we use an operational regulatory region size of 50 kb upstream of the genes transcription start site; a size that has been used in many other studies [27, 91]. In fact, studies have even used regions up to 100 kb [92]. In additionally, we denote the entire region upstream of the gene as the de facto regulatory region of the gene. Together, these assumptions carry the risk that the regulatory region of one gene may contain regulatory sequences for other genes; a more conservative regulatory size would dilute this effect, at the expense of sensitivity. It is not known the extent to which different regulatory sizes and schemes affect the GENMi analysis and more work needs to be conducted on this front. Enhancer promoter interaction data from chromatin capture-based technologies such as Hi-C [93] will help obviate this problem to a certain degree, though such data have to be obtained from the cell type of interest. Furthermore, the GENMi method only considers single TFBS for

filtering eQTL SNPs; associations may be more conspicuous when considering combinations of transcriptional contexts. Though this is hard to compute greedily, there are methods for finding combinations of TFs overrepresented in cis-regulatory regions [94]. Additionally, future work will benefit from analysis of protein QTLsSNPs correlated with protein abundanceas opposed to mRNA abundance [6, 93], as the activity of the gene at the protein level is hypothesized to implement the target cellular response; however, genome-wide protein data are not readily available. Finally, the GENMi method suffers from certain limitations, as seen in Figure 2.4; linkage disequilibrium distributes the effects of casual SNPs in TFBS-promoter interactions among many loci. For regulatory regions with a high incidence of binding and strong cis-eQTL signal, it is likely that our schema for attributing target genes to TFs will induce false positives (incorrect TF-gene interactions). Though at a single gene level this may not be too significant, the GENMi method could still be improved by taking into consideration the local LD environment, as well as the distribution of ChIP tracks in that area. We leave this as future work however.

Our approach utilizing single TF contexts with eQTLs estimated from basal gene expression data in LCLs corresponds to a logical entry point analysis into the pharmacological effects of TFs on drug response. To our knowledge, the GENMi approach is novel in its direct interrogation of transcriptional regulation on drug-induced cellular response. Although many improvements can be made, the fruit of the existing GENMi analysis in both our literature review and control experiments illustrates the remarkable utility of the method.

In the next chapter, we detail the ways in which the GENMi method is extended into a more flexible probabilistic formulation that accommodates multiple types of regulatory evidence.

CHAPTER 3: PGENMI

3.1 INTRODUCTION

In this study, published in Genome Research in 2018 [95], we capitalize on recent trends in genomics to develop a model that integrates multiple types of -omic data to associate TFs with drug response thereby extending the GENMi method from the previous chapter into a more generalized model. Recent studies have analyzed large scale data sets of gene expression to identify genes associated with inter-individual variation in phenotypes ranging from cancer sub-types to drug sensitivity, promising new avenues of research in personalized medicine. However, gene expression data alone is limited in its ability to reveal cis-regulatory mechanisms underlying phenotypic differences. In this study, we develop a new probabilistic model, called pGENMi, that integrates multi-omic data to investigate the transcriptional regulatory mechanisms underlying inter-individual variation of a specific phenotype - that of cell line response to cytotoxic treatment. In particular, pGENMi simultaneously analyzes genotype, DNA methylation, gene expression and transcription factor (TF)-DNA binding data, along with phenotypic measurements, to identify TFs regulating the phenotype. It does so by combining statistical information about expression quantitative trait loci (eQTLs) and expression-correlated methylation marks (eQTMs) located within TF binding sites, as well as observed correlations between gene expression and phenotype variation. Application of pGENMi to data from a panel of lymphoblastoid cell lines treated with 24 drugs, in conjunction with ENCODE TF ChIP data, yielded a number of known as well as novel (TF, Drug) associations. Experimental validations by TF knock-down confirmed 41% of the predicted and tested associations, compared to a 12% confirmation rate of tested non-associations (controls). Extensive literature survey also corroborated 62% of the predicted associations above a stringent threshold. Moreover, associations predicted only when combining eQTL and eQTM data showed higher precision compared to an eQTL-only or eQTM-only analysis using pGENMi, further demonstrating the value of multi-omic integrative analysis.

3.2 METHODS

3.2.1 DATA COLLECTION

We obtained data on single nucleotide polymorphisms (SNPs) , 3 mRNA probe expression, and CpG methylation status across 284 Epstein-Barr Virus (EBV) transformed lym-

phoblastoid cell lines (LCL) from the Coriell Cell Repository. The ethnicities of the cohort decomposed along the following three broad ethnic lines: 95 Han-Chinese American (HCA), 96 Caucasian American (CA), and 93 African American (AA). The sequenced genotype data resulted in 1,362,404 germ line SNPs, each with minor allele frequency (MAF) ¡ 5%, genotype rate $\geq$ 95%, and in Hardy-Weinberg Equilibrium. Imputation analysis of this initial set of genotyped SNPs resulted in 11,256,504 SNPs. Gene expression data consisted of 54,613 Affymetrix U133 Plus 2.0 Gene-Chip probes, transformed using log2 GC Robust Multi-Array Averaging (GC-RMA). Genotyped SNP and gene expression data are available in the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) under SuperSeries accession no. GSE24277 and were originally published in a study by Niu et. al [24]. We used data on 444,797 methylation marks, originally published by Heyn et al [96], available in NCBI GEO under SuperSeries accession no. GSE36369. As a representation of methylation status, we used the beta value, which encodes methylation status within [0, 1], where 0 and 1 correspond to total absence and presence of the mark respectively. Information on pre-processing and analysis of this data including SNP imputation, population stratification, gene expression processing, regression design, among others, resides in further in Methods section for this chapter.

Drug response data was derived from dosage-response curves of 24 cytotoxic drugs: 6-MP, 6-TG, Ara-C, arsenic, carboplatin, cisplatin, cladribine, docetaxel, doxorubicin, epirubicin, everolimus, fludarabine, gemcitabine, hypoxia, metformin, MPA, methotrexate, NAPQI, oxaliplatin, paclitaxel, radiation, rapamycin, triciribine, and temozolomide. Each response curve was summarized by an EC50 value (drug concentration at which half the original LCL population survived treatment). These data were most recently analyzed by Hanson et al. [44] and are available at the following location: veda.cs.uiuc.edu/cytotoxicity. Full information regarding the experimental validation design, data, methodology, and statistical analysis is discussed further in this chapter.

Experimental data on TF binding were obtained from the ENCODE project [1]. The union of Clustered ChIP (v3) tracks of 161 TFs across 91 cell lines formed a single track for each TF, representing a TF composite regulatory profile reflective of activity across a wide variety of cellular contexts. The data for these clustered ENCODE cell lines are located at the following URL: `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTfbsClustered`. We remove clustered peaks that were likely the artifact of high occupancy target (HOT) regions, as in the GENMi analysis [44]. Contrary to that work, we exempted the following 13 TFs from the full 161: general TFs (POLR2A, POLR3A, POLR3G, TBP) and those for which no eQTL or eQTM SNPs ($p < 0.05$) were detected within ChIP peaks (BDP1, BRCA1, BRF1, ELK1, ELK4, ESRRA, HSF1, KDM5A, NELFE).

Further processing of this data, including the removal of High Occupancy Target (HOT) regions, is described in further in this chapter.

3.2.2 TWAS, EQTL, AND EQTM ANALYSIS

Each of the following regression analyses controlled for the following covariates: sex, age, batch, and population axes of variation derived from EIGENSTRAT. Transcript wide association analysis, or TWAS [97], and eQTL analysis were performed following [44]. To perform TWAS, we computed partial regression coefficient p-values between gene expression and drug EC50 values across the 284 LCLs, for each (gene, drug) pair. For eQTL analysis, we calculated partial regression coefficient p-values between each genes expression and the genotype (measured by allelic dosing of 0,1,2) of each SNP in its cis-regulatory (50 Kbp upstream) region. To avoid the statistical artifacts of linkage disequilibrium, we preserved only the most significant eQTL-SNP in each genes cis-regulatory region. For eQTM analysis, we repeated the eQTL analysis while substituting SNP genotypes with methylation status, a continuous variable between 0 and 1 representing the probability of the presence of a CpG methylation, and thus computed eQTM (methylation to gene expression regression) p-values, again retaining only the most significant eQTM in the cis-regulatory region of a gene.

3.2.3 PROBABLISTIC MODEL INTEGRATES GENO/PHENO-TYPE WITH EXPRESSION

OVERVIEW

In our previous work, we presented a proof-of-principle method that identifies TFs associated with individual variation in drug response or any other quantitative phenotype [44].We searched for cases where a SNP in the cis-regulatory region of a gene is correlated with the genes expression (cis-eQTL analysis) and the genes expression is correlated with phenotype - the latter being referred to as a transcriptome wide association study (TWAS) [97]; if significantly many cases like this were identified involving SNPs within ChIP peaks of a TF, then the TF was considered associated with individual variation in phenotype. Here, we constructed a rigorous probabilistic model that builds upon this idea to identify phenotype-related TFs.

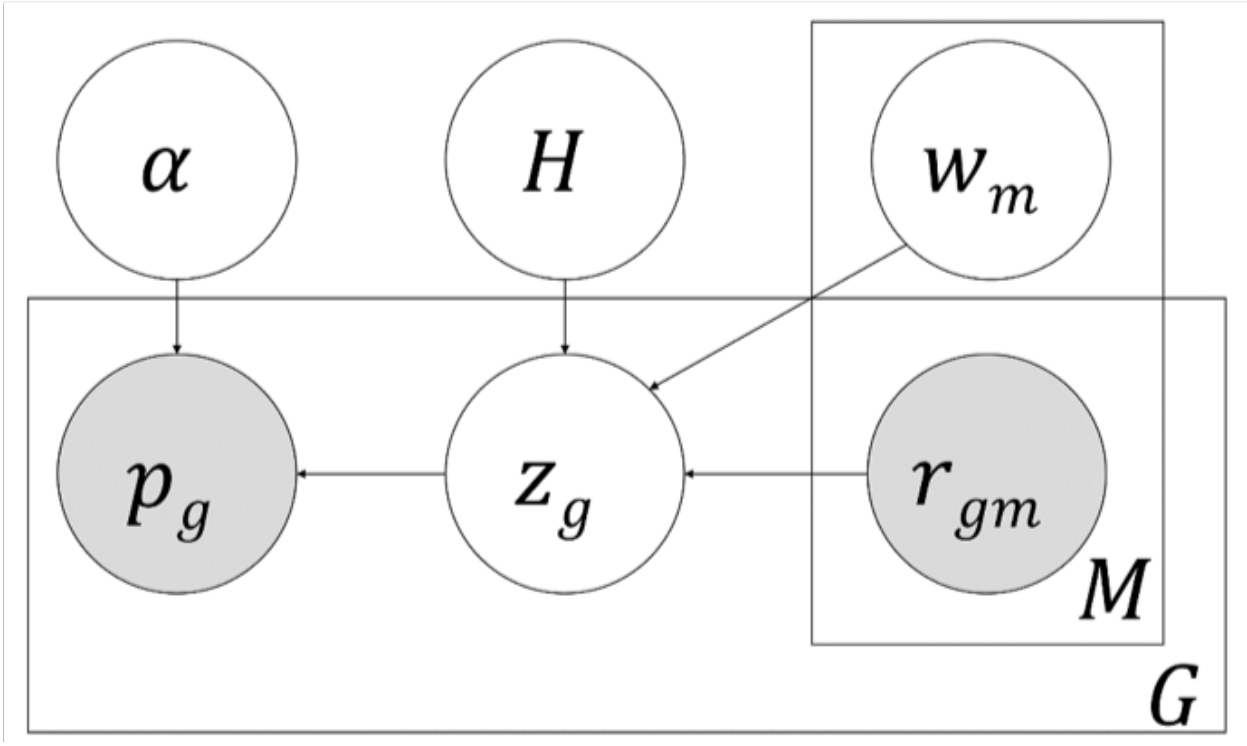


Figure 3.1: Plate diagram of pGENMi model. A latent variable z_g represents whether a gene g mediates the influence of a TF on phenotype, and its enclosing rectangle denotes G such genes. p_g denotes the TWAS p-value between the genes expression and phenotype variation, for or each of the G genes. If $z_g = 1$, we expect an enrichment for significant TWAS p-values, and p_g is modeled by a Beta distribution parameterized by α; otherwise, p_g is modeled as being distributed uniformly in [0,1]. We also observe one or more lines of evidence supporting the TFs influence on the expression of each gene g, such as the existence of a cis-eQTL within a ChIP peak of the TF near that gene. These regulatory evidences are denoted by the binary variables r_{gm}, and there exist M such types of evidence ($m = 1 \ldots M$), with relative weights w_m. These evidences combine in a logistic function to determine $\Pr(z_g = 1)$. The weights w_m are learned over all genes, and as such, are shown outside of the rectangle enclosing G. The H variable indicates whether w_m is free or restricted to 0 (null model), for hypothesis testing.

The new method is called pGENMi (the previous tool was named GENMi for gene expression in the middle and the p denotes a probabilistic model). It integrates information from many genes whose expression correlates with the phenotype, and for which it can find evidence supporting regulatory influence of a specific TF. The probabilistic formulation of pGENMi offers the following important features:

1. It integrates gene expression-phenotype association without relying on strict thresholds.

2. As evidence for a TFs role in gene expression variation, it can utilize information about different types of expression-linked cis-variants (genetic as well as epigenetic) located in the TFs binding sites near the gene.

3. In incorporating multiple sources of evidence for a TFs regulatory role, it can weight the contribution of each type of evidence differently, learning these relative weights automatically.

The probabilistic model of pGENMi is described in Figure 3.1, but we present the main ideas here. The model is evaluated separately for each TF and provides a log likelihood ratio score to quantify the TFs role in regulating phenotypic variation, considering all available data. It assigns to each gene g a hidden variable z_g that takes a value of 1 if the gene g is a mediator of the TFs influence on the phenotype, and 0 otherwise. The case of $z_g = 1$ is supported if the genes expression is correlated with the phenotype (TWAS p-value close to 0), and one or more lines of evidence support the TFs role in regulating that gene. Such regulatory evidence may include the existence of a significant cis-eQTL within the TFs ChIP peak located near the gene. Specifically, the probability of $z_g = 1$ is determined by a linear combination of one or more regulatory evidences, using a single free parameter as the relative weight of each type of evidence. Intuitively, a TF is considered as a potential regulator of the phenotype if the data supports the existence of several genes mediating its influence on the phenotype, as reflected in their z_g variables.

PGM

The pGENMi plate diagram is shown in Figure 3.1, which models the association between a specific TF and a phenotype. There are four variable types $(p, z_g, r_g m, H)$ and two parameter types (α, w_m). The variable $r_g m$ is a binary indicator variable representing whether a gene g has a particular kind of evidence (e.g., cis-eQTL, cis-eQTM, etc.) of regulation by the TF. Here, a cis-eQTL evidence is defined, following[44], as the presence of an eQTL within a ChIP peak of the TF, in the 50Kb upstream region of the gene. Likewise, a cis-eQTM evidence is the presence of an eQTM within a TF ChIP peak in the genes 50 Kbp upstream region. Thus, if $r_{gm} = 1$ for some gene g and m represents cis-eQTM evidence, we interpret this as evidence that a change of TF binding at the ChIP peak is brought about by or related to the observed individual variation in methylation status at the eQTM, which in turn explains the latters correlation with gene expression.

A rectangle labeled with an M engulfs r, indicating that there are M such kinds of regulatory evidence for each gene g: $r_{g1}, r_{g2}, \ldots, r_{gM}$. Each r_{gm} variable in a plate connects to the

same z_g variable. This latter variable is binary and latent (unshaded) and indicates whether expression of a gene g is related to the phenotype. A rectangle labeled G encapsulates z_g and r_{gm}, indicating that there is a latent variable z_g and regulatory evidence vector r_{gm} for each of G genes. The state of z_g is determined probabilistically, as:

$$P(z_g = 1 | r_{g1} \dots r_{gM}, w_0 \dots w_M) = \frac{1}{1 + \exp\left(-(w_0 + w_1 r_{g1} + w_2 r_{g2} + \dots w_M r_{gM})\right)} \tag{3.1}$$

Here, the various evidences r_{gm} for the regulation of gene g by the TF are combined in a logistic function, each weighted by a coefficient w_m. Thus, a weighting of separate regulatory evidences determines whether a gene g is related to the phenotype. The variable p is a continuous variable in the range [0,1], representing the observed TWAS p-value (details above) of the association between gene expression and phenotype. If a gene is related to the phenotype ($z_g = 1$), then we model p to follow a Beta(,1) distribution biased towards small (significant) p-values; is a shape parameter indicating the strength of the bias, constrained to the range [0,1], with =1 equivalent to a uniform distribution over p-values. However, if a gene is unrelated to phenotype ($z_g = 0$), we expect its expression-phenotype correlation p-value to be uniformly distributed over [0,1]. These modeling assumptions are summarized below:

$$p_g \sim \begin{cases} Unif(0,1), & \text{if } z_g = 0 \\ \beta(\alpha, 1), & \text{if } z_g = 1 \end{cases} \tag{3.2}$$

The binary variable H indicates the hypothesis to be tested. When H=1, the model tries to find the best assignments to w and ; this is the alternative hypothesis or H_1. When $H = 0$, the model only trains parameters w_0 and α, and the weights $w_1 \dots w_M$ are removed from the model entirely. In this case, the model tries to explain the observed TWAS p-values without any regulatory evidences at all; this is the null hypothesis, or H_0. We derive a score for how much better the alternative hypothesis explains the data y:

$$LLR = \log_2 P(y|\mathbf{A}, H = 1) - \log_2 P(y|\mathbf{A}, H = 0) \tag{3.3}$$

The pGENMi method uses the log likelihood ratio (LLR), computed for each TF separately, to rank TFs by their predicted association with the phenotype.

PARAMETER ESTIMATION

The model uses the Expectation Maximization algorithm to find assignments to the parameters, $\mathbf{w} = [w_0, w_1, .., w_M]$ and α such that the likelihood of $\mathbf{p} = [p_1, ..., p_G]$ (observed TWAS p-values) is maximized. The optimization imposes no constraints on the parameters, and thus the weights of regulatory evidences ($\mathbf{w}$) may be trained to negative values, in effect rewarding genes that have regulatory evidence but no expression-phenotype correlation; this is inconsistent with our regulatory assumptions, and, as such, we post-process pGENMi results to disregard such spurious TF-phenotype associations.

DETAILED DESCRIPTION

DEFINITIONS M = Number of regulatory evidences + 1.

G = Number of genes.

$\mathbf{R}$ = A matrix of size $G \times M$, where the first columns is all 1s and each additional column a binary corresponding to whether regulatory evidence was observed for that columns source.

$\mathbf{r_g}$:= A vector of length M with the first column as 1 and each additional column a binary corresponding to whether regulatory evidence was observed for that columns source.

$\mathbf{w}$:= A continuous vector of length M that weights each type of regulatory evidence. This is a parameter estimated by the model across all genes.

z_g := A binary latent variable representing whether or not the correlation p-value of gene g's expression with the phenotype of interest is drawn from a uniform or beta distribution.

α := Parameter determining the shape of the beta distribution. The β parameter is set to 1 and α is capped in the range [0,1]. This parameter is estimated by the model.

p_g := A continuous observed variable in the range [0,1] representing the correlation p-value of gene g's expression with the phenotype of interest

LIKELIHOOD

The probability that $z_g = 1$ is a logistic function of $\mathbf{w}$ and $\mathbf{r_g}$:

$$\Pr(z_g = 1|\mathbf{w}, \mathbf{r_g}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})} \tag{3.4}$$

The probability that $z_g = 0$ is 1−probability $z_g = 1$:

$$\Pr(z_g = 1|\mathbf{w}, \mathbf{r_g}) = \frac{\exp(-\mathbf{w} \cdot \mathbf{r_g})}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})} \tag{3.5}$$

If you interpret 1 as $\exp(-0 * r_g)$, then this is simply the Boltzman distribution with the 2nd state being $w = 0$. The probability of p_g depends on if z_g is 1 or 0. With respect to the former, it is distributed per a beta distribution (with $\beta = 1$) and uniform distribution for the latter.

$$\Pr(p_g|z_g = 1, \alpha) \sim \beta(\alpha, 1) \tag{3.6}$$

$$\Pr(p_g|z_g = 0, \alpha) \sim Unif(0, 1) \tag{3.7}$$

Aside:

$$Beta(\alpha, 1) = \frac{x^{\alpha-1}}{\beta(\alpha+1)} = \frac{x^{\alpha-1}}{\Gamma(\alpha)}\Gamma(\alpha+1) = \alpha\frac{\Gamma(\alpha)}{\Gamma(\alpha)}x^{\alpha-1} = \alpha x^{\alpha-1} \tag{3.8}$$

Therefore,

$$\Pr(p_g|z_g = 1, \alpha) \sim \beta(\alpha, 1) = \alpha x^{\alpha-1} \tag{3.9}$$

The likelihood of the data $\mathbf{p}$ given $\mathbf{R}$ is thus.

$$\Pr(\mathbf{p}|\mathbf{w}, \mathbf{R}, \alpha) = \prod_{g=1}^{g} \sum_{i=0}^{1} \Pr(p_g|z_g = i, \alpha)) \Pr(z_g = i|\mathbf{w}, \mathbf{r_g}) \tag{3.10}$$

Expanding the terms and the sum yields the following:

$$\Pr(\mathbf{p}|\mathbf{w}, \mathbf{R}, \alpha) = \prod_{g=1}^{g} \alpha p_g^{\alpha-1} \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})} + \frac{\exp(-\mathbf{w} \cdot \mathbf{r_g})}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})} \tag{3.11}$$

The log likelihood is thus :

$$l(\mathbf{p}|\mathbf{w}, \mathbf{R}, \alpha) = \sum_{i=1}^{g} \log\left(\alpha p_g^{\alpha-1} \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})} + \frac{\exp(-\mathbf{w} \cdot \mathbf{r_g})}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})}\right) \tag{3.12}$$

We can either optimize this function directly or use Expectation Maximization. In this formulation, we choose EM.

POSTERIOR PROBABILITY

The posterior probability for $z_g = 1$ is the following:

$$\begin{aligned} \Pr(z_g = 1|\mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g}) &= \frac{\Pr(z_g = 1, \mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g})}{\Pr(\mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g})} \\ &= \frac{\Pr(z_g = 1, \mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g})}{\Pr(\mathbf{p_g}, z_g = 0, \alpha, \mathbf{w}, \mathbf{r_g}) + \Pr(\mathbf{p_g}, z_g = 1, \alpha, \mathbf{w}, \mathbf{r_g})} \end{aligned} \tag{3.13}$$

$$\Pr(\mathbf{p_g}, z_g = 1, \alpha, \mathbf{w}, \mathbf{r_g}) = \alpha p_g^{\alpha-1} \frac{1}{1 + \exp\left(-\mathbf{w} \cdot \mathbf{r_g}\right)} \tag{3.14}$$

$$\Pr(\mathbf{p_g}, z_g = 0, \alpha, \mathbf{w}, \mathbf{r_g}) = \frac{\exp\left(-\mathbf{w} \cdot \mathbf{r_g}\right)}{1 + \exp(-\mathbf{w} \cdot \mathbf{r_g})} \tag{3.15}$$

$$\Pr(z_g = 1|\mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g}) = \frac{\frac{\alpha p_g^{\alpha-1}}{1+\exp(-\mathbf{w}\cdot\mathbf{r_g})}}{\frac{\alpha p_g^{\alpha-1}}{1+\exp(-\mathbf{w}\cdot\mathbf{r_g})} + \frac{\exp\left(-\mathbf{w}\cdot\mathbf{r_g}\right)}{1+\exp(-\mathbf{w}\cdot\mathbf{r_g})}} \tag{3.16}$$

$$\Pr(z_g = 1|\mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g}) = \frac{\alpha p_g^{\alpha-1}}{\alpha p_g^{\alpha-1} + \exp(-\mathbf{w} \cdot \mathbf{r_g})} \tag{3.17}$$

Naturally, the posterior probability for $z_g = 0$ follows as $1-$posterior probability of $z_g = 1$:

$$\Pr(z_g = 0|\mathbf{p_g}, \alpha, \mathbf{w}, \mathbf{r_g}) = \frac{\exp(-\mathbf{w} \cdot \mathbf{r_g})}{\alpha p_g^{\alpha-1} + \exp(-\mathbf{w} \cdot \mathbf{r_g})} \tag{3.18}$$

EXPECTATION MAXIMIZATION

The function that EM optimizes is the $Q(\theta|\theta^{t-1})$ function where $\theta = w, \alpha$.It is essentially the expectation of the joint log likelihood under the current parameters *theta* at time t with respect to the posterior computed using the previous parameters θ^{t-1}.

$$Q(\theta|\theta^{t-1}) = \sum_{g=1}^{G} \sum_{i=0}^{1} \Pr(z_g = 1|p_g, \alpha^{t-1}, \mathbf{w}^{t-1}, \mathbf{r_g}) \log \Pr(p_g, z_g = i, \alpha, \mathbf{w}, \mathbf{r_g}) \tag{3.19}$$

We will let the posteriors be designated as follows:

$$z_{g0}^{t-1} = \Pr(z_g = 0|p_g, \alpha^{t-1}, \mathbf{w}^{t-1}, \mathbf{r_g}) \tag{3.20}$$

$$z_{g1}^{t-1} = \Pr(z_g = 1|p_g, \alpha^{t-1}, \mathbf{w}^{t-1}, \mathbf{r_g}) \tag{3.21}$$

Expanding terms gives the following:

$$Q(\theta|\theta^{t-1}) = \sum_{g=1}^{G} z_{g1}^{t-1} \log\left(\Pr(p_g|z_g = 1, \alpha) \times \Pr(z_g = 1|\mathbf{w}, \mathbf{r_g})\right) \tag{3.22}$$

$$+ z_{g0}^{t-1} \log\left(\Pr(p_g|z_g = 0, \alpha) \times \Pr(z_g = 0|\mathbf{w}, \mathbf{r_g})\right) \tag{3.23}$$

Further reducing gives:

$$Q(\theta|\theta^{t-1}) = \sum_{g=1}^{G} z_{g1}^{t-1} \log \frac{\alpha p_g^{\alpha-1}}{1+\exp(-\mathbf{w}\cdot\mathbf{r_g})} \tag{3.24}$$

$$+ z_{g0}^{t-1} \log \frac{\exp(-\mathbf{w}\cdot\mathbf{r_g})}{1+\exp(-\mathbf{w}\cdot\mathbf{r_g})} \tag{3.25}$$

Further reducing gives:

$$Q(\theta|\theta^{t-1}) = \sum_{g=1}^{G} z_{g1}^{t-1} \log \alpha p_g^{\alpha-1} - z_{g0}^{t-1}(\mathbf{w}.\mathbf{r_g}) - \log(1+\exp(-\mathbf{w}\cdot\mathbf{r_g})) \tag{3.26}$$

UPDATES

When maximizing, we assume z_{g0}^{t-1} and z_{g0}^{t-1} are fixed. Therefore, finding a formula for $\mathbf{w}$:

$$\frac{\partial Q}{\partial w} = \sum_{g=1}^{G} \mathbf{r_g} * (\frac{\exp(-\mathbf{w}\cdot\mathbf{r_g})}{1+\exp(-\mathbf{w}\cdot\mathbf{r_g})} - z_{g0}^{t-1}) \tag{3.27}$$

Setting to 0 does not yield a closed form solution, so we use this gradient in a gradient descent for updating $\mathbf{w}$.

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \beta \frac{\partial Q}{\partial w} \tag{3.28}$$

For this study, we converge when:

$$||\mathbf{w}^t - \mathbf{w}^{t-1}||_2 < \epsilon \tag{3.29}$$

The parameter α has a closed form solution. Given

$$Q(\theta|\theta^{t-1}) = \sum_{g=1}^{G} z_{g1}^{t-1} \log \alpha p_g^{\alpha-1} - z_{g0}^{t-1}(\mathbf{w}.\mathbf{r_g}) - \log(1+\exp(-\mathbf{w}\cdot\mathbf{r_g})) \tag{3.30}$$

Optimizing for α gives:

$$\frac{\partial Q}{\partial \alpha} = \sum_{g=1}^{G} z_{g1}^{t-1} \frac{\partial}{\partial \alpha}(\log \alpha + (\alpha-1)\log p_g) \tag{3.31}$$

$$\frac{\partial Q}{\partial \alpha} = \sum_{g=1}^{G} z_{g1}^{t-1}(\frac{1}{\alpha} + \log p_g) \tag{3.32}$$

Setting to 0 and solving for α gives the following:

$$0 = \sum_{g=1}^{G} z_{g1}^{t-1}\frac{1}{\alpha} + z_{g1}^{t-1}\log p_g \tag{3.33}$$

$$\frac{-1}{\alpha}\sum_{g=1}^{G} z_{g1}^{t-1} = \sum_{g=1}^{G} z_{g1}^{t-1}\log p_g \tag{3.34}$$

$$\alpha = -\frac{\sum_{g=1}^{G} z_{g1}^{t-1}}{\sum_{g=1}^{G} z_{g1}^{t-1}\log p_g} \tag{3.35}$$

LOG LIKELIHOOD RATIO

We compare the fit model, $H = 1$, to a model where $\mathbf{w} = 0$. We indicate this with, $H = 0$, indicating the null hypothesis. We assess the significance of the alternative model, $H = 1$, by computing the log likelihood ratio between a fit model and the null model.

$$LLR = \log \Pr(\mathbf{p}|\mathbf{w}, \mathbf{R}, \alpha, H = 1) - \log \Pr(\mathbf{p}|\mathbf{w}, \mathbf{R}, \alpha, H = 0) \tag{3.36}$$

IMPLEMENTATION

In our implementation of EM, we choose $\beta = 0.0001$ and $\epsilon = 0.0001$ with a minimum and maximum of 25 and 300 steps in each EM calculation and 100 random restarts for each model. For each random restart, we estimated an alternative and null model. We report the LLR corresponding to the model with the best log-likelihood for the alternative model.

TENSORFLOW IMPLEMENTATION

We also have a TensorFlow implementation available for download, although this implementation was not used to generate the results of this study.

CODE LOCATION

Both the C++ code used to generate the results of this chapter and the TensorFlow code are available in a GitHub repository located at `https://github.com/knoweng/pgenmi` and with a link to the GitHub repository at `veda.cs.uiuc.edu/pgenmi` .

3.2.4 DATA PRE-PROCESSING AND STATISTICAL METHODOLOGY

REMOVING TFS FOR GENMI AND HOT REGIONS FROM ENCODE DATA

Some TFs (ATF1, BCLAF1, MEF2C, PPARGC1A, SMARCC2, STAT2) were filtered out of the GENMi analysis for having an insufficient number of target genes (¡15) for GSEA. We used TF ChIP peaks as a way to focus on SNPs whose association with expression might be mediated by the regulatory action of that TF. However, it is well known that different TFs tend to co-localize at the same genomic loci, a phenomenon that is especially pronounced at HOT (high occupancy target) regions, and a TFs binding at these HOT regions is not necessarily indicative of regulatory function [52]. To enrich the ChIP-based collection of binding sites for functional TF-DNA interactions, we removed segments of 50bp where six or more TFs bind, resulting in a $\sim$25% reduction in the total number of ChIP TF peaks across all ENCODE cell lines.

SNP IMPUTATION

Imputation was run separately for each race and chromosome separately. Chromosomes were divided into 40MB regions. BEAGLE v3.3.1 [98] was run on these 40MB regions of the genome, plus a 1MB buffer region to the right and left of the main region, as the ends are generally imputed poorly. The reference and observed genotype data were input as phased and unphased, respectively. The lowmem option was enforced to reduce the overall amount of memory required to run BEAGLE; additionally, the exclude markers option was used to remove the rare SNPs from the reference mentioned above. For edification, BEAGLE imputed untyped markers, not missing genotyped markers.

SUMMARIZING PROBE EXPRESSION BY GENE SYMBOL

Rather than modeling raw probe data, we opted to summarize probe level expression at the Ensembl gene symbol level, since representing gene expression using multiple correlated probes breaks the independence assumptions of our probabilistic model by giving greater

weight to genes with higher probe coverage. Of the 54,613 probes, those with low variance ($\sigma^2 \leq 0.1$), were omitted from analysis, purging 5,408 probes. Of the remainder, 34,832 probes mapped to 16,183 stable Ensembl gene symbols with at least one exon annotation, using the GRCh37 Ensembl BioMart mapping of Ensembl gene symbols with HGU 133 Plus 2.0 array probe identifiers. Since multiple probes can map to the same gene, we utilized a simple algorithm to obtain a single gene expression value as a function of its mapped probe expressions. The following illustrates this procedure, which uses principle component analysis (PCA) to obtain a vector of gene expression and Z-score normalization to ensure all genes are scaled identically:

For each gene g with at least one mapped probe:

1. If g has one mapped probe
    (a) Return z-score normalization of the probes expression.
2. If gene g has n mapped probes with a probe expression matrix of size n x 284:
    (a) Obtain the first principal component (PC1) of the probe expression matrix.
    (b) Project the probe expression matrix onto PC1.
    (c) Return z-score normalization of the projected PC1 expression.

Applying this procedure resulted in a matrix of Ensembl gene expression with dimensionality 16,183 (number of genes) x 284 (number of LCLs), and Z-score normalized expression across individuals. Henceforth in this publication, we refer to this matrix and its values as the gene expression matrix and gene expression, respectively.

CONTROLLING FOR CONFOUNDING VARIABLES

We included the following potentially confounding variables as covariates in all regression analyses for this study: batch, age, gender, and sub-population labels, derived from EIGENSTRAT (see below). We decided to omit explicit labels of ethnicity, as the sub-population labels should capture that information implicitly. To derive p-values for the covariate of interest in the multiple regression (for instance, a SNP in an eQTL regression) we computed the p-value of the log likelihood ratio between models with and without the covariate of interest, using a χ^2 distribution.

POPULATION STRATIFICATION

Despite the information provided by ethnic labels, it is well known that sub-population structures within these ethnic groups contain important information that can radically confound the results of association analyses when ignored or improperly controlled [86]. To address this issue, we utilized the EIGENSTRAT program (version 6.0.1) developed by the Price lab (`https://www.hsph.harvard.edu/alkes-price/software`), to identify sub-population labels from genotype data and remove their potential confounding effects in regression analyses. Instrumental to EIGENSTRATs ability to recover true population labels is the quality and format in which the data are provided. It may conflate large genomic regions in linkage disequilibrium (LD) with population structure, hence such regions have to be pruned. Additionally, while EIGENSTRAT could identify latent population labels corresponding to the cohorts ethnicities from the entire cohorts genotype matrix, it is more useful run EIGENSTRAT on each known ethnic groups genotype matrix: doing so will direct EIGENSTRAT to look for sub-populations only within each ethnicity, and not across.

To perform this analysis, we first computed a set of independent SNPs using the PLINK program (1.90 beta) [99]. For each ethnicity (HCA, CA, and AA), we used a variant pruning algorithm to remove redundant SNPs in LD and reported only independent SNPs on somatic chromosomes. To generate these results, we ran PLINK with the -indep flag and the following parameters:

1. Window size (SNPs): 50
2. Shift window (SNPs): 5
3. Variance Inflation Factor (VIF) threshold: 1

Using these parameters, PLINK slides a window across the genome and only retains those SNPs in the window that cannot be adequately predicted from a linear combination of the remaining SNPs. For each SNP s in a window, PLINK reports a goodness of fit, R2s, of the multivariate regression to predict s from all other SNPs. PLINK retains only those SNPs with $R^2 >= 1\frac{1}{VIF}$. Setting VIF to 1 ensures that only statistically independent SNPs in each window are selected; although strict, this removes all potential for LD confoundment, at the potential cost of underestimating the number of sub-populations.

Table 3.1: Sub-population principal components extracted from EIGENSTRAT for each ethnicity separately, with Tracy-Widom statistics and p-values.

#	Ethnicity	PC	EigenValue	EigenDiff	TWStat	P-Val	Effect
1	CA	1	1.65	NA	60.84	2E-139	6767.46
2	CA	2	1.48	-1.72E-01	52.05	9.8E-111	9558.58
3	CA	3	1.25	-2.31E-01	18.28	1.58E-24	12486.00
4	CA	4	1.18	-7.04E-02	5.84	2.91E-06	13515.38
5	HCA	1	1.22	NA	9.94	6.74E-11	12980.28
6	AA	1	2.18	NA	73.83	9.9E-186	2955.53
7	AA	2	1.31	-8.68E-01	6.67	3.33E-07	5407.81

We estimated these sub-population labels as continuous axes of variation by using EIGENSTRAT to perform PCA on each ethnic groups independent SNP genotype matrix. Each significant principle component (PC) derived from the genotype matrix of a given ethnicity can be interpreted as a sub-population of that ethnicity. For each ethnicity, we retrieved all PCs with Tracy-Widom p-values ≤ 0.05, resulting in a total of seven axes, shown in Table 3.1. We projected the entire genotype matrix onto each of the seven axes (setting to 0 SNPs not contributing to the axis) to derive numeric sub-population representations for the cohort. We included these sub-population labels as covariates in our regression models to control for population stratification.

REMOVING COMMON TRANSCRIPTION FACTORS FOR EXPERIMENTS

In experimental validation, we avoided TFs like BRF2 and GTF2F1 that were associated with 10 or more drugs. These are general transcription factors whose association is likely due to having many more ChIP peaks than other TFs, a point we have discussed in previous work [44]. We also excluded NAPQI, a toxic byproduct produced during the xenobiotic metabolism of the analgesic paracetamol, due to its lack of clinical application.

CELL CULTURE AND TREATMENTS

All wet-lab experiments described in this chapter were performed in the laboratory of Dr. Liewei Wang at Mayo Clinic, Rochester, MN.

Human triple negative breast cancer (MDA-MB231), leukemia (Jurkat), and glioma (U251) cell lines were obtained from the American Type Culture Collection (Manassas, VA). MDA-

MB-231 cells were cultured in L-15 medium containing 10% Fetal Bovine Serum (FBS). Jurkat cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 Medium, containing 10% FBS. U251 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM), also containing 10% FBS.

The following 10 drugs were purchased from Sigma-Aldrich (St. Louis, MO): 6-Mercaptopurine (6-MP), 6-Thioguanine (6-TG), carboplatin, cisplatin (CDDP), cytarabine (Ara-C), docetaxel, epirubicin, gemcitabine, oxaliplatin, and paclitaxel. The remaining three drugs were obtained Selleck Chemicals (Houston, TX): cladribine, rapamycin (sirolimus), and temozolomide (TMZ). All 13 drugs were dissolved in dimethyl sulfoxide (DMSO) and aliquots of stock solutions were frozen at -80°C.

siRNAs for candidate TFs and negative control siRNA were purchased from Dharmacon. Reverse transfection was performed for MDA-MB231 and U251 cells in 96-well plates. Specifically, 3000-4000 cells were mixed with 0.1 mL of lipofectamine RNAi-MAX reagent (Invitrogen) and 10 nM siRNA for each experiment. Electroporation was performed for Jurkat cells using Nucleofector Kit V from Lonza (Cologne, Germany).

Prior to electroporation, cells were washed with phosphate buffer saline (PBS) and counted. One million Jurkat cells were re-suspended in 100 μL of the Nucleofector Solution buffer and mixed with 100 nM of specific siRNA. The re-suspended cells were transferred to cuvettes and immediately electroporated using the program X-005. After electroporation, cells were incubated in a cuvette at room temperature for 10 minutes and then 500 L of pre-warmed culture medium were added to the cuvette. Cells were then transferred to a 12-well plate and incubated at 37°C/5% CO2 overnight.

Total RNA was isolated from cultured cells transfected with control or TF-specific siRNAs with the Qiagen RNeasy kit (QIAGEN, Inc.), followed by qRT-PCR performed with the one-step Brilliant SYBR Green qRT-PCR master mix kit (Stratagene). Specifically, primers purchased from QIAGEN were used to perform qRT-PCR using the Stratagene Mx3005P Real-Time PCR detection system (Stratagene). All experiments were performed in triplicate with beta-actin as an internal control. Reverse transcribed Universal Human reference RNA (Stratagene) was used to generate a standard curve. Control reactions lacked RNA template.

Table 3.2: Log (base 10) concentration of administered drugs applied during experimental validation (- indicates no drug applied).

Drug	**Log10 Concentration**										**Scale (mol/L)**	
											μ	n
Paclitaxel	-	-2	-1	0	1	1.7	2	3	3.7	4	x	
Docetaxel	-	-2	-1	0	1	1.7	2	3	3.7	4	x	
Epirubicin	-	-1.8	-1.5	-1.2	-0.9	-0.6	-0.3	0	0.3	0.6	x	
Ara-C	-	-5	-4	-3	-2	-1	0	1	2	3	x	
Gemcitabine	-	-5	-4	-3	-2	-1	0	1	2	3	x	
6-MP	-	-2	-1	-0.3	0	0.4	0.7	1	2	3	x	
6-TG	-	-2.3	-1.3	-0.6	-0.3	0	0.4	0.7	1.7	2.7	x	
Carboplatin	-	-4	-3	-2	-1	0	1	2	3	4	x	
Cisplatin	-	0.1	0.4	0.7	1	1.3	1.6	1.8	1.9	2	x	
Oxaliplatin	-	-0.5	-0.2	0.1	0.4	0.7	1	1.3	1.6	1.9	x	
Cladribine	-	-4	-3	-2	-1	0	1	2	3	4		x
Rapamycin	-	-5	-4	-3	-2	-1	0	1	2	3	x	
Temozolomide	-	-0.4	-0.1	0.2	0.5	0.8	1.1	1.4	1.7	2	x	

MTS CYTOTOXICITY ASSAY

Cell proliferation assays were performed in triplicates at each drug concentration. Cytotoxicity assays with the lymphoblastoid and tumor cell lines were performed in triplicates at each dose. Specifically, 90 L of cells (5 104 cells) were plated into 96-well plates (Corning, NY) and were treated with increasing does of a specific drug or radiation. After incubation for 72 hours, 20 L of CellTiter 96®. AQueous Non-Radioactive Cell Proliferation Assay solution (Promega Corporation, Madison, WI) was added to each well. Plates were read in a Safire2 plate reader (Tecan AG, Switzerland).

Cytotoxicity assays with the tumor cell lines were performed using the CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay (Promega Corporation, Madison, WI). Specifically, 90 L of cells (5 103 cells) were plated into 96-well plates and were treated with increasing does of a specific drug. The escalation of log concentrations for each drug is listed in 3.2. After incubation for 72 hours, 20 L of CellTiter 96 AQueous Non-Radioactive Cell Proliferation Assay solution (Promega Corporation, Madison, WI) was added to each well. Plates were read in a Safire2 plate reader (Tecan AG, Switzerland). Cytotoxicity was assessed by plotting cell survival versus drug concentration, on a log scale.

Radiation cytotoxicity was performed in triplicates at each radiation dose as described above. 100 L of cells (5 103 cells) were plated into 96-well plates and were treated with

ionizing radiation at 0, 0.25, 0.5, 1, 2.5, 5, 10, and 20 Gy, using cesium-137 gamma-rays (J.L. Shepherd and Associates Mark I Model 25 Irradiator). After incubation for 72 hours, 20 L of CellTiter 96 AQueous Non-Radioactive Cell Proliferation Assay solution was added to each well. Plates were read in a Safire2 plate reader (Tecan AG).

STATISTICAL ANALYSIS

Significance of the IC50 values between negative control siRNA and TF-specific siRNA was determined using a two-tailed unpaired t-test.

3.2.5 DATA ACCESS

Source code for pGENMi is available at the following:
`https://github.com/knoweng/pgenmi`
`http://veda.cs.uiuc.edu/pGENMi/`

3.3 RESULTS

3.3.1 PGENMI ELUCIDATES ROLE OF TFS IN DRUG RESPONSE VARIATION

We applied the pGENMi algorithm to identify TFs that putatively regulate individual variation in drug response. We considered drug response data of 24 cytotoxic drugs (or treatments) assayed on 284 lymphoblastoid cell lines (LCLs). This phenotypic data was analyzed in conjunction with genotype (SNP), CpG methylation, and gene expression data on the same panel of 284 LCLs, along with ENCODE ChIP-Seq data on TF-DNA binding. To assess the regulatory evidence for a TFs role one a genes expression variation, we first identified the strongest cis-eQTL SNP located near the gene (50Kb upstream of the transcription start site) and within the TFs ChIP peaks; an eQTL p-value 0.05 was considered as evidence for the TFs role in regulating the gene. We similarly tallied eQTM evidence, i.e., a methylation mark significantly correlated with gene expression, located near the gene (cis-eQTM) and within the TFs ChIP peaks. The eQTL and eQTM evidences for the TFs regulatory influence on genes were then integrated with gene expression-phenotype associations (TWAS) using the pGENMi model. The result of this analysis is a ranking (by LLR score) of all TFs by their predicted role in regulating each drug response phenotype. We refer to this mode of analysis as eQTL+M analysis; we also repeated the entire procedure using eQTL-only evidence and eQTM-only evidence

In computing the various correlations (TWAS, eQTL, and eQTM) used in pGENMi analyses, we included as covariates all information for which we had data per individual, i.e., sex, age, and batch, as well as sub-population axes of variation inferred from ethnic labels and genotype information using EIGENSTRAT [89]; Also, we relied on clusters of TF ChIP peaks from many different cell lines, as computed by the ENCODE project, rather than peaks from LCLs exclusively. We believed this would make the regulatory inferences more generalizable across cell lines and the ChIP peaks themselves more likely to be functional. (See Tables 3.10 - 3.11 and Figure 3.4 a comparison to the alternative strategy of using TF peaks exclusively from LCLs).

3.3.2 LITERATURE EVIDENCE

In the absence of a gold standard benchmark, we performed extensive literature search for experimental results corroborating pGENMi-predicted (TF, Drug) associations, using two criteria. The first, and most convincing, is direct experimental evidence demonstrating that knockdown of the TF affects chemosensitivity to the corresponding drug. The second is indirect experimental evidence demonstrating differential expression or DNA binding of the TF induced by the drug.

Table 3.3: Literature support for the (TF, Drug) associations from pGENMi eQTL+M analysis with LLR ≥ 4.5. There are three kinds of literature evidence: desensitization of the cell to the drug as a result of a TF knockdown ("siRNA"), significant change in TF expression ("Diff. Exp.") or binding of the TF to the DNA ("Diff. Binding") upon drug treatment. "None indicates that we failed to find evidence of these three kinds in the literature.

#	TF	Drug	LLR	Literature Evidence	PMID(s)
1	FOXM1	Temozolomide	7.4	siRNA	22977194
2	STAT3	Oxaliplatin	6.95	siRNA	22193989, 21472135, 24098947, 23969971
3	RELA	Temozolomide	6.52	siRNA	23259744, 17638900
4	HNF4G	Epirubicin	6.2	siRNA	26503816
5	HNF4G	Doxorubicin	5.48	siRNA	26503816
6	BATF	Docetaxel	5.57	siRNA	26503816
7	NANOG	Cisplatin	4.62	siRNA	22714588
8	EZH2	Radiation	4.82	siRNA	25460508, 25601206
9	NFIC	NAPQI	7.08	Diff. Expr.	17562736, 21420995, 17585979
10	GATA1	Rapamycin	5.78	Diff. Binding	10713726, 21304100
11	CCNT2	NAPQI	5.67	Diff. Expr.	21420995, 22230336
12	NANOG	Oxaliplatin	5.46	Diff. Expr.	25979230, 26136074, 24462001, 24098947, 23585460
13	ZNF274	Ara-C	5	Diff. Expr.	17039268
14	UBTF	NAPQI	4.8	Diff. Expr.	17562736
15	MEF2C	NAPQI	7.17	None	
16	CTCFL	NAPQI	6.9	None	
17	WRNIP1	6-MP	6.68	None	
18	USF1	Temozolomide	4.69	None	
19	NANOG	6-MP	4.51	None	

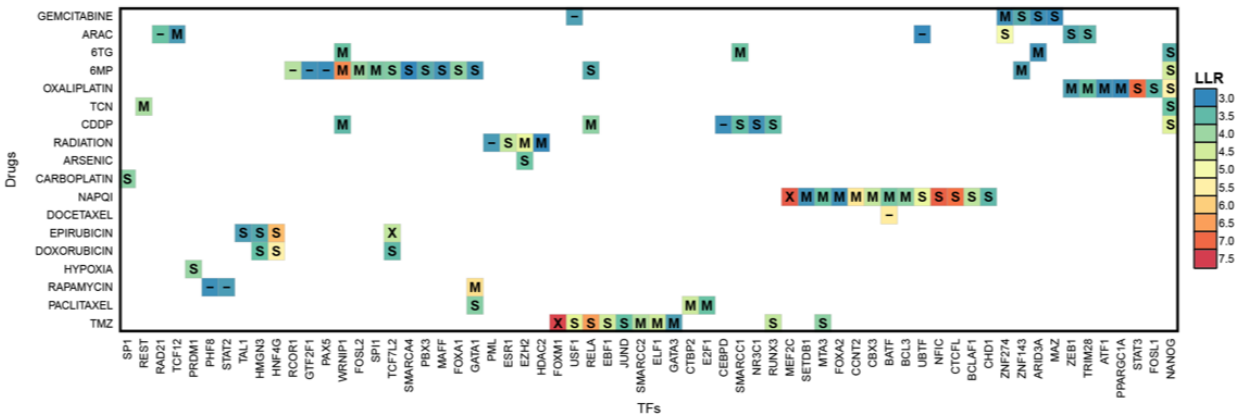


Figure 3.2: The 90 (TF, Drug) associations predicted by eQTL+M with LLR ≥ 3, colored by LLR. Labels indicate whether eQTL-only or eQTM-only analysis corroborated the prediction at LLR ≥ 1.74 (labels S and M respectively). A X indicates both eQTL and eQTM analysis supported the prediction, while a - indicates the prediction is unique to eQTL+M.

Initially, we restricted our literature search to eQTL+M associations with LLR ≥ 4.5, roughly corresponding to a p-value of 0.05 (χ^2 test). Of the 19 reported associations at this stringent criterion, 14 were validated by the literature (8 directly and 6 indirectly), as shown in Table 3.3. For example, pGENMi reported FOXM1 as a top scoring association for the drug Temozolomide; this prediction was validated by a study where siRNA knockdown of FOXM1 was shown to sensitize recurrent glioblastoma multiforme (GBM) tumors to this drug [100]. pGENMi also predicted STAT3 as being associated with response to oxaliplatin. Indeed, siRNA silencing of STAT3 combined with oxaliplatin therapy, in mouse models of metastatic colorectal cancer (HCT116), was previously found to reduce tumor size better than either drug separately [101]. An example of indirect validation is that of the predicted association between GATA1 and rapamycin. Treatment of hexamethylenebisacetamide (HMBA), which commits cells to cessation of growth and differentiation, to friend erythroleukaemia cells increased DNA binding of GATA1, an important TF for erythroid specific genes. When treated with the S6-kinase inhibitor, rapamycin, HMBA cells induced at 18 hours showed markedly lower binding of GATA1 to the DNA [102].

Some drug responses may be easier to model using gene expression than others, due to variation in the quality of experiments and different mechanisms of action. As such, we sought to examine how literature-based validations of pGENMi associations segregate by drug. To do this, we relaxed the LLR threshold to 3, which resulted in 90 (TF, Drug) associations (Figure 3.2).

Table 3.4: Literature support for predicted (TF, Drug) associations, separated by drug. Shown are the number of predictions (P), true positive (TP), and precision of up to top 7 associated TFs with LLR *geq*3 for each drug. We call a predicted (TF, Drug) pair true if there is direct or indirect literature evidence for it.

Drug	**P**	**TP**		**Precision**	
		Direct	**Indirect**		
6-MP	7	0	1	1/7	14%
6-TG	4	0	0	0/4	0%
Ara-C	6	3	2	5/6	83%
Arsenic	1	1	0	1/1	100%
Carboplatin	1	1	0	1/1	100%
Cisplatin	7	5	2	7/7	100%
Docetaxel	1	1	0	1/1	100%
Doxorubicin	3	2	1	3/3	100%
Epirubicin	4	3	0	3/4	75%
Gemcitabine	5	0	1	1/5	20%
Hypoxia	1	1	0	1/1	100%
NAPQI	7	0	4	4/7	57%
Oxaliplatin	7	2	4	6/7	86%
Paclitaxel	3	1	1	2/3	67%
Radiation	4	4	0	4/4	100%
Rapamycin	3	3	0	3/3	100%
TCN	2	0	0	0/2	0%
Temozolomide	7	3	0	3/7	43%

We further restricted our literature survey to at most top seven associations for each drug (to limit the burden of manual validation), amounting to 73 associations. The precision (fraction of positive predictions supported by literature evidence) of these top associations varies widely across drugs (Table 3.4). For certain drugs, such as cisplatin, all seven predictions were confirmed by direct or indirect literature evidence. We also observed 100% precision for the following drugs for which a single TF was predicted:. On the other hand, TFs associated with certain drugs, e.g., 6-MP and 6-TG, were rarely supported by the literature. Overall, our literature survey resulted in validation of 45 of 73 (62%) predictions made at the modest threshold of LLR ≥ 3, but pointed to substantial inter-drug variability in pGENMis ability to identify TF determinants of cytotoxicity.

Table 3.5: Decomposition of the 90 (TF, Drug) associations with LLR >= 3 from eQTL+M analysis by their significance in eQTL- and eQTM-only analyses.

		eQTM-only	
		No	Yes
eQTL	No	11	31
	Yes	45	3

Table 3.6: Decomposition of the LLR ranks of eQTL and eQTM (TF, Drug) associations with no thresholds applied.

A		**eQTM**	
		Top 500	Not Top
eQTL	Top 500	103	384
	Not Top	397	2595

(a) Decomposition of the LLR ranks of eQTL and eQTM (TF, Drug) associations with no thresholds applied.

B		**eQTM+M**	
		Top 500	Not Top
eQTL	Top 500	309	161
	Not Top	173	2765

(b) Decomposition of the LLR ranks of eQTL and eQTL+M (TF, Drug) associations with no thresholds applied.

C		**eQTM+M**	
		Top 500	Not Top
eQTL	Top 500	262	223
	Not Top	220	2703

(c) Decomposition of the LLR ranks of eQTM and eQTL+M (TF, Drug) associations with no thresholds applied.

We also assessed the benefit of modeling multiple types of regulatory evidence simultaneously (i.e., eQTL+M) rather than simply taking the union of associations reported by running pGENMi for each evidence (eQTL or eQTM) separately. To answer this question, we ran eQTL-only and eQTM-only analyses and applied an LLR threshold of 1.74 (roughly corresponding to χ^2 p-value of 0.05) for their reported associations. We then categorized the 90 eQTL+M associations with LLR$\geq$ 3 (Figure 3.2) based on their recapitulation in eQTL- and eQTM-only analysis (Table 3.5). Strikingly, the eQTL+M associations were rarely supported by both analysis. Despite this, 11 associations were identified by the eQTL+M model that a simple intersection would not produce; we further investigated these associations in our experimental validation. In looking at the top 500 (TF, drug) associations reported by each analysis, we noted that the eQTL+M analysis and eQTL-only analysis showed strong concordance (309 common associations), while the eQTM-only analysis showed much less concordance with results from the other two analyses (see Table 3.6)

TOP 19 (TF, DRUG) ASSOCIATIONS

Here, we list and discuss the top (TF, Drug) associations predicted by eQTL+M with an LLR $\geq$ 4.5 ranked by strength of evidence and LLR score (see Table 3.3). For each (TF, Drug) association, we report several fields. The first is the "Validation Status" field, which indicates whether we consider this association to be direct validation (exhibited usually by TF knockdown) or indirect (exhibited usually by TF mRNA differential expression or binding in response to the drug treatment). The second field, "Evidence", summarizes the evidence we found for the association. The next field, "LLR", contains the pGENMi eQTL+M LLR for the association. The following High Scoring Analyses field contains the experiments that yielded a significant LLR; for eQTL and eQTM, the LLR threshold was 1.74. Since we only examine significant eQTL+M pairs in this analysis, every association should at least have eQTL+M in this field. The last field is Observation which is a paragraph describing literature evidence we found corroborating the (TF, Drug) association or at least providing some evidence as to its validity.

Association: FOXM1 with Temozolomide
Validation Status: Direct
Evidence: TF siRNA and Differential TF Concentration
LLR: 7.40
High Scoring Analyses: eQTL+M, eQTL, eQTM
Observation: Recurrent glioblastoma multiforme (GBM) tumors are characterized by a resistance to chemo and radiotherapy. Recurrent GBM tumors resistant to temozolomide showed

higher levels of FOXM1 than primary tumors. Treating GBM cell lines with temozolomide increased expression of FOXM1 and the DNA damage repair gene, RAD51 and showed resistance to temozolomide treatment. Knockdown of FOXM1 via siRNA assays inhibited RAD51 expression and sensitized recurrent GBM to temozolomide. The regulatory relationship between FOXM1 and RAD51 was further corroborated by ChIP analysis showing a preponderance of FOXM1 binding in the RAD51 promoter [100].

Association: MEF2C with
Validation Status: Not Validated (Some Evidence)
Evidence: Regulatory Partner of Differentially Expressed Drug Target mRNA
LLR: 7.17
High Scoring Analyses: eQTL+M, eQTL, eQTM
Observation: Yeast 2 hybrid experiments showed an interaction between MEF2C and HABP4; further investigation showed that HABP4 inhibits the DNA binding potential of MEF2C. This interaction was confirmed in-vitro using GST-pull down assays and in-vivo rat heart cells by ChIP [103]. A separate study on hepatotoxicity demonstrated that acetaminophen treatment affected HABP4 expression [104]. Given the interaction between MEF2C and HABP4 and HABP4s sensitivity to acetaminophen, we concluded that though there may be an association between MEF2C and NAPQI, we did not have sufficient experimental evidence to call the relationship direct or indirect.

Association: NFIC, CCNT2, and UBTF with NAPQI
Validation Status: Indirect
Evidence: Differential Expression of TF mRNA
LLR: 7.08, 5.67, and 4.80 respectively
High Scoring Analyses: eQTL+M for all, eQTL for NFIC, and eQTM for CCNT2 and UBTF
Observation: A multi-center study of the effect of acetaminophen (APAP) toxicity to liver cells showed differential expression of the TFs NFIC and UBTF to acetaminophen treatment [108]. In a separate gene expression analysis of human liver slices, treatment of acetaminophen also induced differential expression NFIC and CCNT2 [105]. The association between NFIC and APAP toxicity was also established in mouse liver samples [106]. Finally, a transcriptomic, proteomic, and metabolomics profile of hepatoma cells with and without APAP demonstrated the differential CCNT2 expression [107].

Association: STAT3 with Oxaliplatin
Validation Status: Direct
Evidence: TF siRNA and Differential TF Concentration
LLR: 6.95
High Scoring Analyses: eQTL+M, eQTL

Observation: In a study exploring the immune response of oxaliplatin treatment, researchers showed that oxaliplatin decreased TLR-induced STAT1 and STAT3 expression in human T cells via Western Blotting [108]. In mouse models of metastatic colorectal cancer (HCT116), siRNA silencing of STAT3 combined with oxaliplatin therapy reduced tumor size by 96%, better than either treatment separately (77% and 57% respectively) [101]. This interaction between oxaliplatin and STAT3 was articulated in another study of HCT116 cells, where in vitro treatment of oxaliplatin was accompanied with IL-6 mediated activation of STAT3 as well as phosphorylation of Raf kinase inhibitor protein (RKIP) [109]. Work by Hua et al. in SKOV3 ovarian cancer cell lines recapitulated this relationship in a different cell line, with the conclusion being that oxaliplatin treatment upregulated STAT3β [110].

Association: RELA with Temozolomide

Validation Status: Direct

Evidence: TF siRNA and Differential Expression of TF mRNA Expression

LLR: 6.52

High Scoring Analyses: eQTL+M, eQTL

Observation: A previous study demonstrated that the DNA - methylating drug temozolomide (temozolomide) activates the positive modulator of NF - B, AKT, in a mismatch repair (MMR) system dependent manner. Mismatch repair systems are either proficient in the repair of DNA double strands breaks (DSBs), or deficient, conferring chemosensitivity to the former and chemoresistance to the latter. A subsequent investigation into whether NF B is activated by temozolomide and whether AKT is involved in the molecular biology of this event revealed several interactions between the NF B family member, RELA, and temozolomide. Treatment of temozolomide to proficient MMR systems enhanced NF B transcriptional activity, activated AKT, and induced RELA nuclear translocation in only MMR-proficient cells. Upregulation of NF B transcription and RELA translocation were impaired in KD12 cells treated with temozolomide and transfected with siRNA targeting AKT. Additionally, RELA silencing in deficient MMR systems increased temozolomide-induced growth suppression [111]. An examination of temozolomide inhibition of NF B activity revealed that O6-methylguanine inhibits RELA DNA binding; Another study showed differential RELA expression in response to temozolomide treatment in glioma cells [112].

Association: HNF4G with Epirubicin and Doxorubicin

Validation Status: Direct

Evidence: TF siRNA

LLR: 6.20 and 5.48 respectively

High Scoring Analyses: eQTL+M and eQTL for both

Observation: Our previous work demonstrated the link between the anthracyclines, epiru-

bicin and doxorubicin, with HNF4G by using siRNA knockdown of HNF4G in two triple negative breast cancer cell lines MDA-MB-231 and BTF459. The cytotoxicity curve of cell population survivability versus the log concentration of epirubicin or doxorubicin treated was significantly altered in both MDA-MB-231 and BTF459 when transfected with HNF4G siRNA [44]: and compared to DMSO control experiments.

Association: GATA1 and Rapamycin

Validation Status: Direct

Evidence: Differential TF DNA Binding and siRNA

LLR: 5.78 *High Scoring Analyses*: eQTL+M, eQTM *Observation*: There is a close relationship between the mTOR pathway, which rapamycin directly targets, and the PI3K share a lot of crosstalk, leading to the development of dual inhibitors [113]. Research on the involvement of phosphoinosite 3-kinases (PI3K) in cellular differentiation was investigated in the context of friend murine erythroleukaemia cells. The early hours of dimethyl sulfoxide (DMSO) or hexamethylenebisacetamide (HMBA) exposure to these cells commits them to a cessation of growth and differentiation. Treatment of these inducers to friend erythroleukaemia cells increased DNA binding of GATA1, an important transcription factor for erythroid specific genes. When treated with the S6-kinase inhibitor, rapamycin, HMBA cells induced at 18 hours showed markedly lower binding of GATA1 to the DNA [106]. Together, this indicated that rapamycin may inhibit binding of GATA1 to DNA via the PI3K dependent AKT/p70 S6-kinase pathway. Another study concluded that mammalian target of rapamycin (mTOR), which is related to the aforementioned AKT pathway, tightly regulates GATA1 protein expression at the post-transcriptional level [114]. Yet another investigation into the role of epidermal growth factor (EGF) on upregulation of the excision-repair cross-complementary 1 (ERCC1) gene in human hepatocarcinoma cells (HuH7) revealed that in EGF-induced HuH7 cells, PI3K inhibition combined with silencing of the PI3K pathway kinase FKBP12-rapamycin-associated protein or mammalian target of rapamycin (FRAP/mTOR) upregulated ERCC1. Additionally, motif search identified a binding site for GATA1 in the promoter of ERCC1 with ChIP confirming the binding of GATA1 to the promoter in EGF-induced cells [115]. This indicates that GATA1 plays a role in mTOR/PI3K signaling, as it is a possible regulator of a target of such a pathway. Finally, this association was also corroborated by our own siRNA knockdown of GATA1 in MDA-MB-231 cells treated with rapamycin; cells were more resistant to rapamycin-induced apoptosis when GATA1 was knocked down.

Association: BATF with Docetaxel

Validation Status: Direct

Evidence: TF siRNA

LLR: 5.57

High Scoring Analyses: eQTL+M

Observation: In the study by Hanson et al., the effect of BATF on the cytotoxicity of paclitaxel and docetaxel on the triple negative breast cancer cell lines, BTF549 and MDA-MB-231. Silencing BATF shifted both the cytotoxicity curves of paclitaxel and docetaxel in MDA-MB-231 significantly with respect to DMSO controls, but not in BTF549 [44].

Association: NANOG with Oxaliplatin

Validation Status: Indirect

Evidence: Differential Expression of TF mRNA and Differential TF Protein Concentration

LLR: 5.46

High Scoring Analyses: eQTM only

Observation: Ex vivo models derived from human colorectal liver metastases were treated with oxaliplatin, 5-flurouracil, and curcumin. Compared to DMSO controls, at 72 hours of treatment, the triplicate treatment significantly downregulated expression of NANOG [116]. NANOG protein concentration was also shown to be decreased in colorectal cancer stem cells (CRSCs) when treated with thiostrepton, which acts synergistically with oxaliplatin in killing CRSCs [117]. Additionally, research has shown that NANOG amplifies STAT3 activation, a regulator shown to be associated with oxaliplatin cellular response [118, 109]. Finally, the Notch signaling pathway, which targets NANOG [123], was shown to be an important mediator of CRSC self-renewal and proliferation (`http://ecommons.luc.edu/luc_diss/87`).

Association: ZNF274 with Ara-C

Validation Status: Indirect

Evidence: Differential Expression of TF mRNA

LLR: 5.00

High Scoring Analyses: eQTL+M, eQTL

Observation: While no literature evidence was found that corroborates this finding, evidence does exist for the association between gemcitabine, which, like Ara-C, is a deoxycytidine analog. The link between gemcitabine and ZNF274 was reported by eQTL+M analysis with an LLR of 3.11. A gene expression profile of gemcitabine treatment to breast cancer cells revealed that gemcitabine upregulates ZNF274 expression [119].

Association: EZH2 with Radiation

Validation Status: Direct

Evidence: TF siRNA, Differential Expression of TF mRNA and Regulatory Information

LLR: 4.83

High Scoring Analyses: eQTL+M, eQTM

Observation: In a study of epigenetic aberration in radioresistant OML1-R cells, researchers

identified hypermethylation of the FHIT promoter by H3K27me3 and low FHIT expression. Further analysis revealed EZH2 overexpression in these cells. EZH2 is a known transcriptional silencer via H3K27me3 and so the investigators knocked down EZH2 via shRNA to assess its impact on FHIT. Knockdown of EZH2 increased FHIT expression, decreased H3K27me3 in the FHIT promoter by 2-fold, increased H3K4me3 in the FHIT promoter by 2-fold, and reduced FHIT promoter methylation by 10%. An inhibitor of EZH2 H3K27me3, GSK343, resulted in increased expression of FHIT; combination therapy with the DNA methyltransferase inhibitor (DNMT) 5-Aza demonstrated the greatest increase of FHIT expression among all epigenetic silencers [120]. In a study of glioblastoma (GBM) derived tumorigenic stem-like cells (GSCs), Kim et al. looked at the role that the catalytic subunit of Polycomb repressive complex 2, EZH2, and the MELK-FOXM1 complex play in radiosensitivity. It was shown that not only are both EZH2 and MELK co-expressed in GBM and upregulated following radiation treatment, but that MELK mediated EZH2 signaling is required for GSC resistance to radiation. They further show that this function is evolutionarily conserved in Caenorhabditis elegans (C. elegans). The researchers further detailed the mechanisms by which EZH2, MELK, and FOXM1 may be interacting. Luciferase assays in overexpression and knockout experiments showed increased EZH2 promoter activity in the presence of MELK; this finding was corroborated by flow cytometry experiments. The conclusion of these findings was that MELK transcriptionally regulated EZH2, at least in GBM spheres. Given the lack of a MELK DNA binding domain, the researchers searched for a cofactor that MELK teamed with to regulate the expression of EZH2. The paper concluded that EZH2 is a direct target of the MELK/FOXM1 transcriptional complex [121].

Association: NANOG with Cisplatin.

Validation Status: Direct

Evidence: TF siRNA

LLR: 4.62

High Scoring Analyses: eQTL+M, eQTL

Observation: The role of NANOG in esophageal squamous cell carcinoma (ESCC) was investigated, particularly its interaction with cisplatin. ESCCs were confirmed to have high expression of NANOG via RT-PCR. Several other squamous cell lines have been characterized by NANOG expression, which was hypothesized to be the reason for these cell lines chemosensitivity to cisplatin. When NANOG was silenced via siRNA in two ESCC cell lines, the growth inhibitory effect of cisplatin was significantly enhanced [122].

Table 3.7: All 73 (TF, Drug) Associations in the Top 7 associations for each drug (LLR >3). D.E indicates Differential Expression of mRNA and D.C. indicates Differential Concentration of Protein. The subject/object of each experiment is presumed to be the TF unless otherwise stated.

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
1	WRNIP1			6.68	X	X	
2	NANOG			4.51	X		X
3	RCOR1	Indirect	Reg by Drug Target	4.23	X		
4	FOSL2			3.99	X	X	
5	FOXA1			3.89	X		X
6	TCF7L2			3.77	X		X
7	SPI1			3.74	X	X	

(a) 6-MP

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
8	SMARCC1			7.26	X	X	
9	WRNIP1			7.20	X	X	
10	NANOG	Indirect		6.81	X		X
11	ARID3A			6.24	X	X	

(b) 6-TG

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
12	ZNF274	Indirect	D.E.,	5.00	X		X
13	RAD21	Direct	siRNA and Cancer Remission	3.59	X		
14	TRIM28	Indirect	D.P. (Gemcitabine)	3.44	X		X
15	ZEB1	Direct	Knockdown	3.31	X		X
16	TCF12	Indirect	D.E.	3.20	X	X	
17	UBTF			3.03	X		

(c) Ara-C

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
18	EZH2	Direct	Kinase siRNA and miRNA siRNA and Overexpression	3.62	X	X	

(d) Arsenic

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
19	SP1	Direct	ChIP and O.E.	3.85	X		X

(e) Carboplatin

Table 3.7: continued ...

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
20	NANOG	Direct	siRNA	4.62	X		X
21	RELA	Direct	Reporter Assay and siRNA	3.88	X	X	
22	SMARCC1	Direct	Indirect shRNA and Homolog deletion	3.53	X		X
23	RUNX3	Direct	siRNA and D.E.	3.51	X	X	
24	WRNIP1	Indirect	Cooperators siRNA	3.49	X	X	
25	NR3C1	Indirect	D.E.	3.10	X	X	
26	CEBPD	Direct	Transfection and D.E./D.C.	3.07	X		

(f) Cisplatin

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
27	BATF	Direct	siRNA	5.57	X		

(g) Docetaxel

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
28	HNF4G	Direct	siRNA	5.48	X		X
29	HMGN3	Direct	siRNA	3.55	X		X
30	TCF7L2	Indirect	Pathway Evidence and D.C.	3.50	X		X

(h) Doxorubicin

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
31	HNF4G	Direct	siRNA	5.48	X		X
32	HMGN3	Direct	siRNA (Doxorubicin)	3.55	X		X
33	TCF7L2			3.50	X		X
34	TAL1	Direct	TF Overexpression	3.21	X		X

(i) Epirubicin

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
35	ZNF143	Some	D.E. and TF siRNA (Cisplatin)	3.37	X		X
36	USF1	Some	D.E. Drug Target and Regulatory Info	3.24	X		
37	ZNF274	Indirect	D.E.	3.11	X	X	
38	ARID3A			3.06	X		X
39	MAZ			3.03	X		X

(j) Gemcitabine

Table 3.7: continued ...

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
40	PRDM1	Direct	Reg. Drug Target and D.E.	3.96	X		X

(k) Hypoxia

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
41	MEF2C	Some	Reg. Partner of D.E. Drug Target	7.17	X	X	X
42	NFIC	Indirect	D.E.	7.08	X		X
43	CTCFL			6.90	X		X
44	CCNT2	Indirect	D.E.	5.67	X	X	
45	UBTF	Indirect	D.E.	4.90	X		X
46	CBX3			4.37	X	X	
47	BCLAF1	Indirect	D.C.	4.31	X		X

(l) NAPQI

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
48	STAT3	Direct	siRNA and D.C.	6.95	X		X
49	NANOG	Indirect	D.E. and D.C.	5.46	X		X
50	TRIM28	Indirect	D.P and D.C.	3.58	X	X	
51	FOSL1	Indirect	D.E. and Network Analysis	3.47	X		X
52	ZEB1	Indirect	Reg. Drug Regulator	3.26	X	X	
53	ATF1			3.10	X	X	
54	PPARGC1A	Direct	TF siRNA	3.05	X	X	

(m) Oxaliplatin

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
55	CTBP2			4.44	X	X	
56	GATA1	Indirect	D.C. and Genomic Alteration	4.81	X		X
57	E2F1	Direct	D.C. and O.E.	3.58	X	X	

(n) Paclitaxel

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
58	EZH2	Direct	siRNA, D.E., and Reg. Info	4.83	X	X	
59	ESR1	Direct	TF Activation and D.E.	4.36	X		X
60	PML1	Direct	Fluorescence, siRNa, O.E., and N Blot	3.19	X		
61	HDAC2	Direct	D.C. and siRNA	3.02	X	X	

(o) Radiation

Table 3.7: continued ...

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
62	GATA1	Direct	D.B. and siRNA	5.78	X	X	
63	STAT2	Direct	Fluorescence, siRNa, and W Blot	3.20	X		
64	PHF8	Direct	O.E. and siRNA	3.08	X		

(p) Rapamycin

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
65	REST	Indirect	Reg. Drug Target	4.14	X	X	
66	NANOG	Indirect	Reg by Drug Target	3.60	X		X

(q) Triciribine

#	TF	Evidence		LLR	eQTL ___		
		Code	Type		L+M	M	L
67	FOXM1	Direct	siRNA and D.C.	7.40	X	X	X
68	RELA	Direct	siRNA and D.E.	6.52	X		X
69	USF1			4.69	X		X
70	EBF1	Some		4.49	X		X
71	ELF1	Direct	siRNA	4.46	X	X	
72	RUNX3			4.35	X		X
73	SMARCC2			4.23	X	X	

(r) Temozolomide

Table 3.8: Both tables show experimentally tested (TF, Drug) pairs, selected based on their high LLR scores, and with trained parameter values either suggesting an effect of the TF on drug (A) or no effect (B). For each predicted pair, shown is the LLR from the eQTL+M analysis and whether there was support (marked as x) from eQTL, eQTM, and eQTL+M analyses. The Validated column and grey shading shows whether the experimental test found a significant effect on cytotoxicity upon knockdown of TF. Table (B) reveals experimental validation of only 1 of 8 tested pairs, consistent with our expectation that these pairs should prove false despite a high LLR, based on parameter values.

Drug	TF	LLR	eQTL	eQTM	eQTL+M	Validated
Epirubicin	TCF7L2	8.70	x	x	x	No
Temozolomide	FOXM1	14.81	x	x	x	No
6-MP	WRNIP1	13.03		x	x	No
Temozolomide	SMARCC2	8.45		x	x	No
Rapamycin	GATA1	10.00		x	x	Yes
Temozolomide	ELF1	8.53		x	x	Yes
6-MP	RCOR1	8.46			x	No
6-MP	PAX5	6.13			x	No
Ara-C	UBTF	6.06			x	No
Gemcitabine	USF1	6.48			x	No
Radiation	PML	6.37			x	No
Ara-C	RAD21	7.17			x	Yes
Cisplatin	CEBPD	6.14			x	Yes
Docetaxel	BATF	11.14			x	Yes
Rapamycin	STAT2	6.41			x	Yes
Rapamycin	PHF8	6.16			x	Yes
Temozolomide	ZNF263	5.55		x		No

(a) Experimentally tested (TF, Drug) pairs in the predicted group, organized by support from eQTL, eQTM, and eQTL+M analyses.

Drug	TF	LLR	eQTL	eQTM	eQTL+M	Validated
6-MP	EZH2	19.98	x	x	x	Yes
6-TG	EZH2	23.88	x	x	x	No
6-TG	FOXP2	11.30	x	x	x	No
Carboplatin	TBL1XR1	9.27	x	x	x	No
Paclitaxel	CHD1	8.21	x	x	x	No
6-MP	FOXP2	6.44			x	No
Oxaliplatin	NR3C1	6.64			x	No
Cisplatin	EZH2	8.87		x	x	No

(b) Experimentally tested (TF, Drug) pairs in negative control group of associations, where parameter values suggest lack of evidence for TF association, despite a high LLR score.

TOP 73 (TF, DRUG) ASSOCIATIONS BY DRUG

Here, we list and discuss the top (TF, Drug) associations predicted by eQTL+M with an LLR$\geq$ 3, decomposed by treatment, and ranked by LLR. For each treatment, only up to seven TFs were examined (see Table 3.7). In parenthesis is the fraction of pairs validated for each drug, where validation is either direct (knockdown or overexpression) or indirect (differential expression, differential binding, etc). Observations are also provided for pairs where we observed some evidence, but not enough to be considered either direct or indirect validation.

For each drug, we attempted to validate in literature (also leveraging experimental data from our own analyses on a portion of association) the top 7 associations with an LLR$>$ 3, resulting in 73 (TF, Drug) associations (see Table 3.7). For those with some, indirect, or direct evidence of association, we summarize the evidence alongside citations to each of the referred papers in Appendix A.

3.3.3 EXPERIMENTAL VALIDATION

Certain associations we wanted to test ourselves in the laboratory. We sought to experimentally verify whether TFs associated with drug response variation (by pGENMi) could be linked in vivo to significant changes in drug-induced cytotoxicity.

We selected predicted (TF, Drug) pairs that reflect a diversity of regulatory support, shown in Table 3.8. For instance, we selected 10 of the 11 (TF, Drug) pairs identified uniquely by eQTL+M analysis (we omitted GTF2F1), and two of the three (TF, Drug) pairs predicted by all three analyses: eQTL+M, eQTL, and eQTM. Additionally, we selected the top four pairs by LLR in the eQTL+M analysis, that were also supported by eQTM-only analysis. We also included a high-scoring association from eQTM-only analysis that scored poorly in the eQTL+M analysis (i.e., LLR$<$ 3).

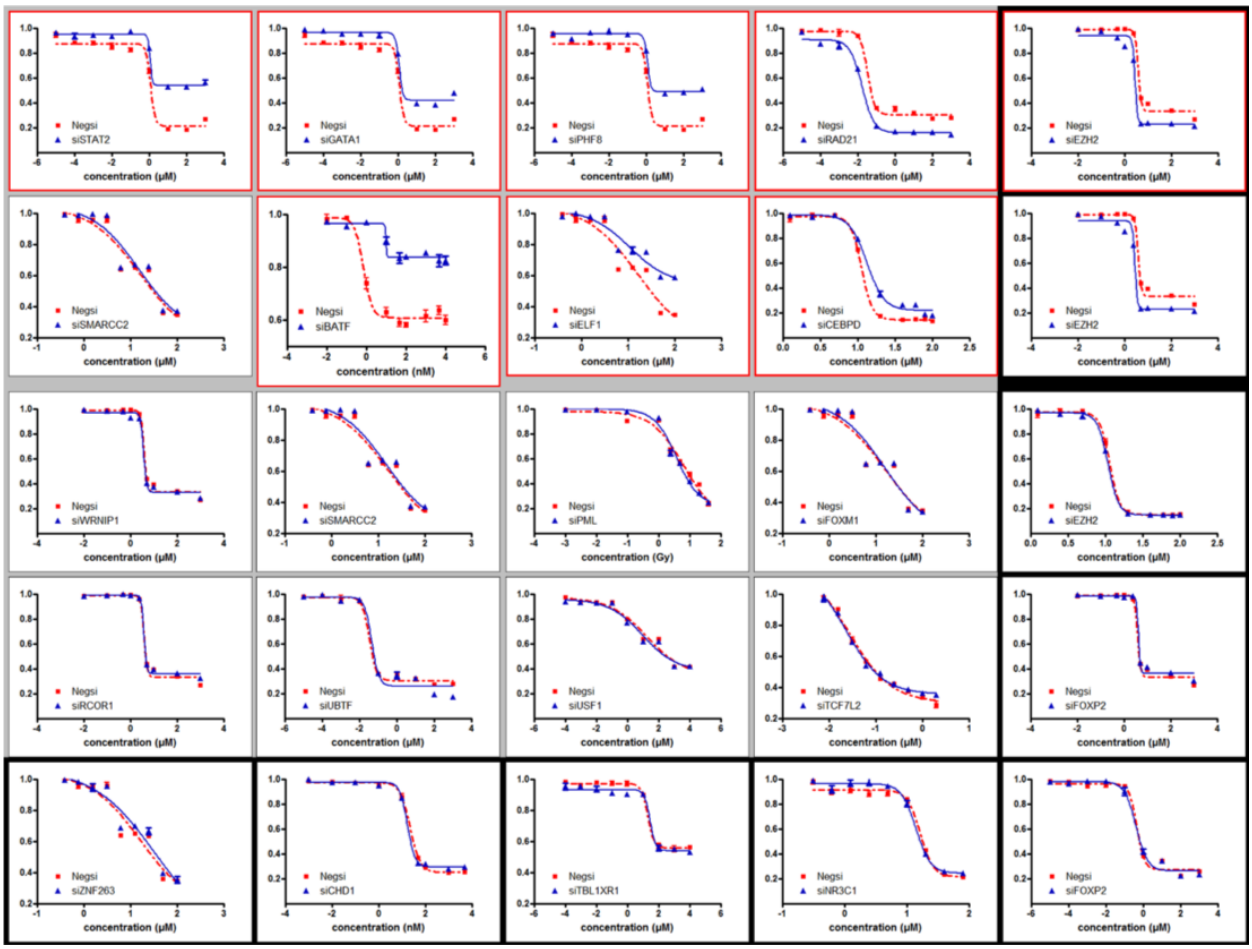


Figure 3.3: All 25 experimentally tested (TF, Drug) dosage-response curves. Red outlines indicate significant shifts in cytotoxicity between siRNA negative and siRNA TF conditions. Curves with a grey background are eQTL+M predictions, while those with a black background are negative controls. We confirmed 7 out of 17 predictions and only 1 out of eight negative controls.

We did not experimentally pursue eQTL-only predictions, as such validations were reported in our previous work [44] and our focus here was primarily on the eQTL+M mode. All 17 of the these (TF, Drug) pairs were found by pGENMi using learnt parameters consistent with our expectation, i.e., where presence of regulatory evidence relating a TF to a gene makes the gene more likely to be associated with the drug (parameters $w_i > 0$, for $i \neq 0$). For some (TF, Drug) pairs, the maximum likelihood estimation assigned negative values to both of the w_i parameters, while yielding a high LLR score, marking a departure from the null model. However, such negative weights indicate that the absence of regulatory evidence relating the TF to the gene is a better marker for linking genes to phenotype, meaning that

the TF is unlikely to be a regulator of the phenotype. We included eight such (TF, Drug) associations as a negative control group, pushing the total number of experimental pairs to 25. We expected these eight predictions to prove false, since pGENMi considered them as statistically interesting but in a manner inconsistent with its mechanistic underpinnings.

Though we largely utilized lymphoblastoid cell lines data for our association analysis, we performed siRNA knockdown experiments in several different cell lines to demonstrate the generalizability of our results beyond LCLs. Based on clinical relevance, the human triple negative breast cancer MDA-MB-231 cells were chosen to test anthracyclines, taxanes, platinums, gemcitabine, radiation, and rapamycin cytotoxicity, because these drugs are typically administered as first-line therapy for triple negative breast cancer. We used a human leukemia Jurkat cell line to test 6-MP, 6-TG, and Ara-C since these drugs are used to treat leukemia. Temozolomide is the first-line therapy for glioblastoma multiforme; we therefore used human glioma U251 cells to validate the TFs associated with Temozolomide response. The siRNA knockdowns were performed for the 25 (TF, Drug) pairs shown in Table 3.8 for the eQTL+M predicted and negative control associations respectively. Cytotoxicity graphs of all knockdowns are shown in Figures 3.3, while those separated by drug are shown in Appendix B. Overall, 7 of 17 ($\sim$41%) of predicted associations (Table 3.8A), were validated in this way, compared to 1 of 8 ($\sim$ 12%) of negative control associations (Table 3.8B). Interestingly, this overall precision of 41% in our own validations is similar to the corresponding precision observed across all drugs (Table 3.5, Direct evidence). Additionally, the precision on the 10 unique eQTL+M predicted associations was 50% (5 of 10), compared to 0% (0 of 2) of negative control associations unique to eQTL+M mode; this indicates pGENMi is able to combine multiple lines of regulatory evidence to make novel regulatory predictions that may be unreported when considering each line of evidence in isolation. With respect to the (TF, Drug) pair predicted uniquely from eQTM analysis (and not predicted eQTL- or eQTLM-only), we failed to validate it in our experiments. Nevertheless, we believe that our experimental validation demonstrates the utility of pGENMi predictions overall and especially when combining multiple sources of regulatory information to learn novel associations.

3.3.4 PGENMI PRODUCES MORE DISTINCT ASSOCIATIONS THAN SIMPLE METHODS

The pGENMi eQTL+M analysis, with results restricted to those with an LLR$\geq$ 3, produced 90 (TF, Drug) associations. As no such algorithm other than GENMi [44] exists for associating TF cis-regulatory influence with drug response, and even GENMi is incapable of

handling both genotype and methylation information simultaneously, there were no obvious reported baseline methods to which we could compare pGENMi so we constructed two simple baselines. In the first, a TF is considered associated with a drug if the TFs expression correlates significantly with cytotoxicity, an approach recently used for identifying genes associated with drug response [36]. This baseline approach reported 121 associations at FDR 0.1 (chosen so that the number of reported associations is similar to that of pGENMi). However, these associations were mostly distinct from those reported by pGENMi; between these 121 associations and the 90 associations reported by pGENMi, only five associations (hypergeometric p-value = 0.19) were shared. Similar observations of complementarity between results were made when using other thresholds of significance.

In a second baseline method, a TF is associated with a drug if the top drug response genes are enriched for strong nearby ChIP peaks of that TF. Performing a hypergeometric test between the top 500 target genes of a TF, based on the maximum ChIP score in the 50 Kb upstream region of each gene, and the top 500 drug response genes from a TWAS analysis yielded 126 (TF, Drug) pairs at an FDR of 0.1 (correction performed per drug). Similar to the first baseline, only two were shared with the pGENMi approach. The results of both baseline analyses demonstrate that pGENMi reveals novel insights into regulatory mechanisms of drug response, based on cis-regulatory analysis, that may not be obtained from the current approach of expression-phenotype correlations or from analyses that identify gene targets purely on the basis of the distribution of strong ChIP peaks.

We also investigated the extent to which sub-significant results in the baseline were enriched with pGENMi. Relaxing the 0.10 FDR threshold to 0.25 produced 253 associations rather than 121, though the overlap with the 90 pGENMi associations only increased to 6 (hypergeometric p-value = 0.63). We further diluted the statistical reliability of the TWAS associations by reporting associations based on uncorrected p-values ≤ 0.05. Of the 384 associations matching this criterion (among the 3552 test), the overlap with pGENMi yielded only 10 associations (hypergeometric p-value = 0.51).

3.3.5 TOP PGENMI ASSOCIATIONS OVERLAP WITH GENMI BUT MOST ARE NOVEL

pGENMi formalizes the statistical procedure of GENMi [44] and extends it to handle multiple regulatory evidences. As such, it is fair to ask to what extent these two related procedures produce similar results. We therefore ran GENMi to test the enrichment between genes ranked by their TWAS p-value pertaining to a given drug and genes with either cis-eQTL or cis-eQTM evidence for that TF.

Table 3.9: Concordance between pGENMi and GENMi eQTL+M and eQTL-only models. Intersection/Union (Jaccard Coefficient) of top 90 (TF, Drug) associations from various pairs of analyses that differ in model (pGENMi versus GENMi) or data (eQTL+M versus eQTL-only).

Top 90 (TF, Drug) pairs		pGENMi		GENMi	
		eQTL+M	eQTL-only	eQTL+M	eQTL-only
pGENMi	eQTL+M				
	eQTL-only	48/132 (36%)			
GENMi	eQTL+M	24/156 (15%)	24/156 (15%)		
	eQTL-only	22/158 (14%)	29/151 (19%)	44/136(32%)	

Comparing the top 90 (TF, Drug) pairs by p-value produced by GENMi and the top 90 (TF, Drug) pairs produced by pGENMi using both eQTL and eQTM evidence reveals an overlap of 24 (TF, Drug) pairs (hypergeometric test p-value 1.1 E-17), corresponding to a Jaccard coefficient of 15.4%, shown in Table 3.9. Of these 24 pairs, ten showed directly validation either through our experiments or in the literature. While a strong overlap between the top associations from these two related methods is expected, it also underscores that pGENMi finds several associations 66 of the top 90 that were not reported by GENMi using the same rank threshold (47 were ranked below 300, out of 3000, by GENMi). This complementarity arises due to the difference of models used in the two analyses, since the same data were utilized.

Comparison of GENMi and pGENMi results (top 90 pairs) obtained from eQTL-only or eQTL+M data, in all four combinations, revealed an overlap of 22-29 pairs (Table 3.9). The commonalities were much greater when results from eQTL+M analysis were compared to those from eQTL-only analysis, using the same method, than when pGENMi and GENMi were compared using the same data (Table 3.9). For example, the top 90 predictions by pGENMi with eQTL+eQTM data and by pGENMi with eQTL data share 48 associations, while top 90 predictions by pGENMi with eQTL data and by GENMi with eQTL data share only 29 associations. Thus, roughly speaking, a greater difference in predicted associations arises in changing the model (from GENMi to pGENMi) than in adding eQTM data to pGENMi analysis (from pGENMi eQTL-only to pGENMi eQTL+M).

In the absence of a reliable benchmark of true associations and given the impracticality of extensive literature-based assessment of each of the analyses reported in Table 3.9, it is difficult to argue empirically if one analysis is more accurate than another. However, it is fair to conclude that the majority of the top associations reported by pGENMi are exclusive

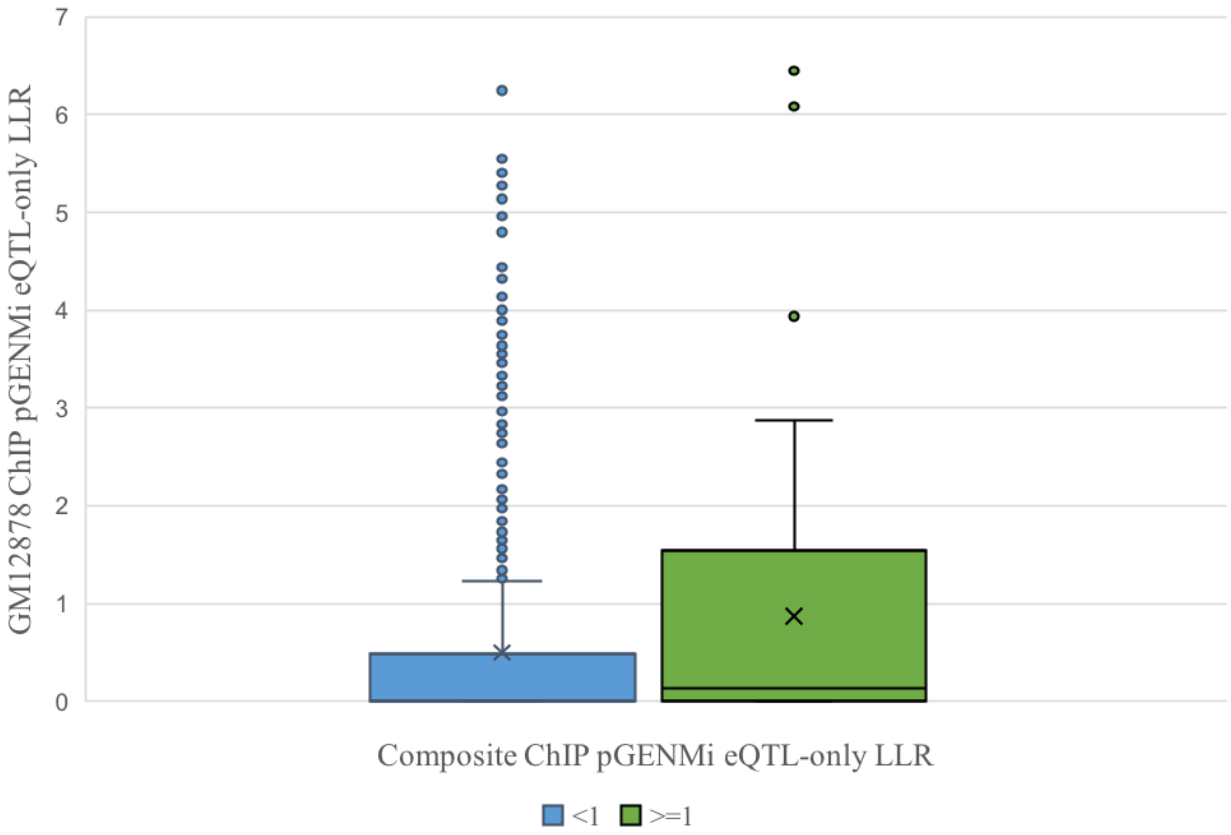


Figure 3.4: Distribution of GM12878 ChIP pGENMi eQTL-only Log-Likelihood Ratios (LLR) for tested (TF, drug) pairs for which the corresponding LLR with Composite ChIP data is < 1 (blue) or ≥ 1(green) respectively.

to it in comparison with results from GENMi, regardless of the mode of analysis.

3.3.6 CELL LINE SPECIFIC CHIP DATA AND COMPOSITE CHIP DATA ALTER RESULTS

As the genotype, methylation, gene expression, and cytotoxicity data all were derived from the same panel of cell lines, we inquired whether restricting ChIP peaks to those from lymphoblastoid cell lines produced different results than when using a union of clustered peaks across cell lines (composite ChIP data), which was the strategy used for the analyses above. To test this, we ran pGENMi using eQTL data co-incident with GM12878 ChIP peaks from ENCODE and compared the results to the eQTL-only analysis using composite ChIP peaks.

Among the 37 GM12878 specific TFs and 24 drugs tested, we tested 888 (TF, Drug)

associations, of which 90 pairs exhibited an LLR ≥ 1.74. At the same threshold, and when testing the same candidate pairs, the composite ChIP analysis produced 29 associations, with only three associations (ELF1, 6-MP), (MAX, 6-MP), and (NFIC, 6-TG) being common with the analysis based on GM12878 peaks. We noted that pairs with LLR ≥ 1 in composite ChIP analysis had a greater tendency (compared with those with LLR < 1) to be in the significant list of the GM12878 ChIP analysis (t-test p-value 3.05E-04, see Figure 3.4) Overall, our conclusion from this comparison is that the top associations reported by pGENMi depend significantly on the source of the ChIP data use to infer regulatory evidence, which is expected. Our use of composite (multiple cell line) ChIP data in the results reported in this work was motivated by our search for associations that generalize beyond lymphoblastoid cell lines, to other cancer cell lines where many of the associations validated in the literature as well in this study have been demonstrated.

Table 3.10: The decomposition of TFs associated with drugs among the top 90 (TF, Drug) pairs for the GM12878 eQTL-only pGENMi analysis

Drug	**# TFs associated in top 90 (TF, Drug) pairs**
Rapamycin	1
Triciribine	1
Hypoxia	2
NAPQI	2
Gemcitabine	2
Fludarabine	4
Radiation	5
Arsenic	6
6-MP	14
Doxorubicin	16
Epirubicin	18
6-TG	19

It is notable that the GM12878 specific analyses produced a substantially greater number of associations above the reporting threshold. One reason for this may lie in the difference between the GM12878 and Composite ChIP analyses in terms of the number of target genes (those with regulatory evidence) for each TF. For instance, the median number of target genes with cis-eQTL evidence for a TFs influence is 58 per TF and 151 per TF for Composite and GM12878 ChIP data respectively; the mean number of target genes per TF based on either cis-eQTL or cis-eQTM (eQTL+M) evidence improves to 102 for the composite ChIP data and 155 for the GM12878 data.

Table 3.11: The decomposition of drugs associated with TFs among the top 90 (TF, Drug) pairs for the GM12878 eQTL-only pGENMi analysis

Drug	**# TFs associated in top 90 (TF, Drug) pairs**
BHLHE40	1
EGR1	1
BATF	1
RELA	1
USF2	1
SPI1	1
BCL11A	1
USF1	1
NFYB	2
EBF1	2
PAX5	2
ZEB1	2
POU2F2	2
NFATC1	2
IRF4	2
MEF2A	2
BCL3	2
ATF2	2
TCF3	2
RUNX3	3
TCF12	3
SRF	3
MTA3	3
ZNF143	4
MAX	4
NFIC	4
MXI1	5
FOXM1	5
YY1	5
ELF1	5
SP1	5
SIN3A	5
PML	6

Table 3.12: The number of TWAS genes (of 16183) at varying p-value thresholds and the number of pGENMi eQTL+M associations for those drugs at LLR ≥ 3

Drug	**# TWAS gene with p-value $\leq$**		**pGENMi eQTL+M associations at LLR ≥ 3**
	0.005	**0.0005**	
6-MP	956	397	15
NAPQI	103	12	13
Temozolomide	324	133	10
Cisplatin	862	265	7
Oxaliplatin	562	227	7
Cytarabine	854	376	6
Gemcitabine	298	68	5
6-TG	882	377	4
Epirubicin	313	64	4
Radiation	181	27	4
Doxorubicin	318	74	3
Paclitaxel	176	34	3
Rapamycin	113	21	3
Triciribine	257	77	2
Arsenic	147	22	1
Carboplatin	1339	489	1
Docetaxel	138	27	1
Hypoxia	82	6	1
Cladribine	61	6	0
Everolimus	79	17	0
Fludarabine	88	12	0
Metformin	368	57	0
MPA	112	9	0
Methotrexate	56	7	0

The detection discrepancy between the two analyses that leads to fewer targets in the Composite ChIP analysis may be due, in part, to the clustering across cell lines which dilutes LCL specific ChIP signals and high occupancy target (HOT) region removal. It does not seem exceptional that LCL derived eQTLs would be more enriched in LCL specific ChIP peaks. Another reason for the GM12878 specific analyses exhibiting may be the asymmetric distribution of TFs associated with drugs in the analysis using GM12878 ChIP data.

As shown in Table 3.10, the 4 drugs 6-MP, Doxorubicin, Epirubicin, 6-TG account for 67 of the 90 (TF, Drug) associations observed at a LLR ≥ 1.74. Two of these drugs 6-MP

and 6-TG also had among the largest number of TWAS genes, which together with the larger numbers of target genes designated with GM12878 ChIP data may have resulted in the inflated numbers of TF associations. The enrichment of associations for these drugs may also indicate a greater sensitivity to them in this cell line. Table 3.11 shows the distribution of associated drugs for each TF at the LLR ≥ 1.74 threshold.

We maintain that the composite ChIP analysis offers a more balanced and generalizable view of the regulatory landscape for a TF and thus is more robust to validation in other cell lines and tissues. We leave more in depth analysis of the utility of cell line specific ChIP data for future research.

3.3.7 POTENTIAL FACTORS LEADING TO FALSE POSITIVES IN PGENMI

Validation rates of pGENMi predictions, based on literature search or our own experiments, were around 40% overall, which is far from perfect, and highly variable across drugs. One of the potential reasons for false positives is the great variation among drugs in terms of the number of significant TWAS genes, i.e., genes with significant correlation between their expression and cytotoxicity.

Table 3.13: The median and mean number of TWAS genes (of 16183) at varying p-value thresholds and the median and mean number of pGENMi eQTL+M associations at LLR ≥ 3 for the Top 5 and Bottom 19 drugs by the # of pGENMi eQTL+M associations at ≥ 3

Statistics		**# TWAS gene with p-value $\leq$**		**pGENMi eQTL+M associations at LLR ≥ 3**
		0.005	**0.0005**	
Top 5 Drugs	Median	562	227	10
	Mean	561.4	206.8	10.4
Bottom 19 Drugs	Median	176	27	1
	Mean	308.5	93.2	2

We tabulated the number of significant TWAS genes alongside the number of pGENMi eQTL+M associations ($LLR \geq 3$) for each drug in Table 3.12. The five drugs with most reported associations (6-MP, NAPQI, Temozolomide, Cisplatin, Ara-C) - about 10 TFs on average - show far more TWAS genes (median of 227), compared to the remaining 19 drugs, which typically yield about 1 TF association (median) and have only 27 TWAS genes

(median); see Table 3.13. This raises the possibility that pGENMi makes false positive predictions in cases when there are significantly more phenotype-associated genes. Pursuing this further, we noted that for the top five drugs by number of associations, our literature-based validation (Table 3.4) found direct evidence for only 10 of 35 (29%) predicted TF associations, while for the remaining 19 drugs we were able to find similar direct evidence for 20 of 38 (53%) TF associations predicted by pGENMi, a statistically significant difference (hypergeometric test p-value 0.032). Similar observations were made on our own experimental validations (Table 3.8A), though small sample sizes prevent statistical claims. For instance, two of the above-mentioned top 5 drugs had 3 or more TFs tested by us (6-MP and Temozolomide), and both yielded low validation rates (0/3 and 1/4 respectively), while the drug Rapamycin, for which only 3 predictions were made and all three were tested by us, yielded a success rate of 3/3. In light of the above observations, we believe that the accuracy of pGENMi predictions may suffer when a phenotype variation is correlated with a large number of genes expression levels, and future work should attempt to rectify this issue if possible.

Another possible factor leading to false positives is that extensive co-binding of a pair of TFs leads to the method mistaking the co-bound TF for the true regulator. We have tried to address this potential source of false positives by removing HOT regions where most co-binding is observed from our analysis. Many TF pairs are known to exhibit significant co-binding even after HOT region removal, as demonstrated by the ENCODE project, and thus it is possible that some of the false positives arise from co-binding. In such cases, if the true TF is among those tested by us (i.e., has ENCODE ChIP data), then we would expect that it should also score highly in the pGENMi analysis. For example, in our experimental validation, we predicted that ELF1 and ZNF263 were both associated with temozolomide, but only succeeded in validating ELF1 through cytotoxicity assays, while ZNF263 knockdown did not show significant change in response. ENCODE reports that these two TFs co-bind across the whole genome in K562 cells [123], so it is possible that the predicted ZNF263 association was a false positive because of the true ELF1 association and the ELF1-ZNF263 co-binding.

Finally, with regards to the experimental validation, some (TF, Drug) pairs didnt validate in vivo because of the assay design; (FOXM1, Temozolomide) and (PML, Radiation) were validated directly in tissues according to the literature, but both failed to validate in vivo in cell lines. Furthermore, we validated (FOXM1, Temozolomide) in recurrent glioblastoma multiforme (GBM) tumor but failed to validate in U251 which is derived from completely different tissue than GBM. The same holds for (PML, Radiation), which was validated most strongly in mouse embryonic fibroblasts (MEFs), and which failed in a triple negative breast

cancer cell line (MDA-MB-231) completely unlike MEFs. Because (TF, Drug) pairs may be context (cell-line) specific, failure to validate in certain cell lines cannot be interpreted necessarily as a rejection of the relationship altogether, as the two examples above show.

3.3.8 PGENMI COMPARISON TO GPA

The Genomic Pleiotropy with Annotation (GPA) [88] algorithm is similar in spirit to pGENMi. GPA uses latent variables for SNPs and integrates the GWAS p-value of a SNP with annotations of that SNP; pGENMi in contrast uses latent variables for genes and integrates the TWAS p-value a gene with annotations of that gene based on eQTL, eQTM and TF ChIP peak evidence. The commonality between the two models, is thus at a higher conceptual level rather than at a practical level that might warrant empirical comparisons. Even at the technical level there is a key difference in that GPA models the joint likelihood of the annotations and GWAS p-value of SNPs, whereas we only use regulatory evidence to maximize the likelihood of TWAS data.

3.3.9 DATABASE OF (TF, DRUG) ASSOCIATIONS

We made available as an online resource all (TF, Drug) associations validated using siRNA or overexpression experiments in this study as well as those found to be similarly validated in our survey of the literature, the GENMi study [44], and a related work that performs the same experimental validations [124]. This resource is available at veda.cs.uiuc.edu/pgenmi. We believe this to be the first such catalog of experimentally validated (TF, Drug) pairs.

CHAPTER 4: CONCLUSION

In this work, we have demonstrated the ability to identify TFs which work through intermediary target genes to implement the cellular response to drugs, via a statistical enrichment model (GENMi). We also extended such a statistical model into a more flexible probabilistic formulation that can successfully associate TFs with drug response via multiple regulatory modalities, in a manner that generates novel associations that would not have otherwise been discernable from either analysis individually. Both of these methods provide principled ways to discover the regulatory mechanisms underlying phenotypic variation among individuals. We demonstrated a high rate of success in validating our associations through a comprehensive literature survey and in vivo experimental validation and have collated the validations from our and other studies into a (TF, Drug) association database.

While we view the pGENMi model as a major improvement over our previous proof-of-principle approach [44], it is similar to various existing models for data integration. pGENMi somewhat resembles the GPA model [88] except that GPA, by modeling SNPs instead of genes as latent variables, applies primarily to GWAS data analysis. Furthermore, we use an empirical Bayesian approach similarly used to prioritize SNPs using annotations [91] and Battle et. als latent regulatory variant model (LRVM) [92]. The pGENMi model, although unique in the way that it combines cis-regulatory information with inter-individual co-variation in gene expression and drug response, is not wholly unique in identifying TFs relevant to phenotype. There exist categories of algorithms dedicated to implicating transcriptional mechanisms or regulatory networks specific to certain phenotypes [125] [126] [127]. The main distinction of the pGENMi approach is that it uses data on population-level variations in genotype, epigenotype, gene expression and phenotype to uncover those mechanisms. Its findings are thus expected to be more relevant to what underlies individual differences in phenotype, as opposed to a more general catalog that includes TFs that may be relevant to the phenotype but not so much to its individual-level variations. On a relatively minor note, the above methods for reconstructing context-specific transcriptional regulatory networks require data on discrete cellular conditions (disease vs non-disease state for instance) and are thus distinct from pGENMi which focuses on inter-individual continuous variation in cellular phenotype. TF associations with drugs can also be inferred from general studies that prioritize genes related to drug response based on prior functional networks [124] [128] [129] or based on observed or imputed gene expression alone [130] [36] [33] [131]. These approaches are not focused on identifying regulators of phenotype based on

cis-regulatory evidence.

Validation rates of pGENMi predictions, based on literature search or our own experiments, were around 40% overall, which is far from perfect, and highly variable across drugs. Exploring possible reasons for false positives, we noted great disparity among drugs in terms of the number of significant TWAS genes (i.e., whose expression is associated with phenotypic variation), and that pGENMi predicts more associations, with lower validation rates, for drugs with more TWAS genes (Table 3.12-3.13). In some cases, extensive co-binding of a pair of TFs may also have led to false positives, with peaks of the true regulator mistakenly providing evidence in favor of the co-binding TFs influence on several TWAS genes. With regards to experimental validation, some (TF, Drug) pairs didnt validate in vivo because of the assay design; (FOXM1, Temozolomide) and (PML, Radiation) were validated directly in tissues according to the literature, but both failed to validate in vivo in cell lines. Because (TF, Drug) pairs may be context (cell-line) specific, failure to validate in certain cell lines cannot be interpreted necessarily as a rejection of the relationship altogether.

pGENMi can be improved in a number of ways; for instance, our model makes the implicit assumption that a latent variable zg representing a genes involvement in mediating the TFs influence on phenotype is independent across genes; however, co-regulated gene sets break this assumption and it would be more reflective of the underlying biology to model a network of latent variables zg instead of treating each gene independently. Our choice of modeling the p-values of gene expression-phenotype correlation as arising from a Beta distribution avoids the use of arbitrary thresholds on the strength of this relationship, but we believe there may be less restrictive ways to explicitly model phenotype-expression relationships. Although we integrated more types of data in this work than typical studies that directly relate gene expression to drug response data [36] [130]), the model allows further extensions. For instance, incorporating the impact of various histone modifications and other epigenetic marks will be an interesting endeavor if such variants are profiled in the cell line panel. Additionally, it may be wise to consider in vivo assays to validate the mediators of drug response, in addition to the TFs themselves, in the future. Despite these various areas of improvement, the pGENMi methodology was successful in integrating both genetic and epigenetic variation to elucidate the regulatory association between a TF and drug, as evidenced by our 50% precision in validating those predictions only eQTL+M reported. We hope this methodology can serve as a blueprint for future endeavors that aim to elucidate the regulatory basis of disease etiology using multiple molecular profiles of variation.

The pGENMi framework need not be restricted to the analysis of the variation of drug response. It may be easily adapted for use in other situations where one seeks to link a phenotype to a set of gene properties, e.g., regulation by a TF, membership in a pathway, involvement in a biological process, etc., while using gene expression as an intermediate variable. The relationship between gene expression and phenotype may be quantified by differential expression p-values if, for instance, the phenotype is binary as in case versus control studies or in before versus after treatment studies.

From a computational modeling standpoint, one might wonder about the utility of choosing this pGENMi probabilistic methodology; after all, there are many different categories of machine learning algorithms from which one might choose. We concede that the point of this endeavor was not to survey the relevant and applicable methods and to identify the best such one for this task; the task is novel. Therefore, our focus was to identify a model that could answer the biological questions we set forth. The probabilistic methodology we choose is natural in the sense that (a) it is a probabilistic extension of the intuitive statistical methodology GENMi and (b) similar methods have been published analyzing different entities such as GWAS data, albeit with slightly different graphical structure. We understand the limitations in using summary data (p-values) rather than relying on the samples themselves in the analysis. However, the works outlined in this paper are association studies and, therefore, it is appropriate to use summary statistics, especially when data comes from disparate sources. In moving forward, it will be instrumental to incorporate a sample level framework in moving from association analysis to analyses that can make stronger claims.

Regarding the problem as a whole of identifying regulatory mechanisms underlying drug response variation, there are number of ways in which this work can be extended. Moving away from treating each target gene independently from other genes to a network-based approach is one such logical extension; analyzing regulatory and drug response variation in this context empowers clinical and laboratory scientists, as much of the literature describing drug response variation is on the level of pathways rather than individual genes. Another avenue lies in integrating the entire signal transduction pathway, incorporating phosphorylation and other signaling data in addition to the -omic correlations in this study. This extension would help filter out spurious correlations at the -omic level while boosting marginally significant correlations whose relationships are consistent with the underlying signaling. Incorporating prior biological knowledge into these analyses would help tremendously. Thus far, we have relied on relations between -omics entities derived from the dataset itself. However, any single dataset is likely to be biased and incomplete and incorporating both known biological

associations and de novo associations will likely merge the best of both worlds. Finally, the drug response models and TF-target models used in this study are simplistic. More elaborate drug response models and TF-target gene models that incorporate both expression and ChIP data will obviously go a long way toward improving the general algorithmic performance. Overall, the work this thesis represents is an entry point into the intersection of pharmaco- and regulatory genomics and their remains many future work to fully analyze the relationship between these two genomic domains.

APPENDIX A: TOP 73 (TF, DRUG) ASSOCIATION VALIDATIONS

This chapter contains the literature research conducted for those TFs whose validation in Table 3.7 was not mentioned. For each (TF, Drug) association, we report several fields. The first is the Validation Status field, which indicates whether we consider this association to be direct validation (exhibited usually by TF knockdown) or indirect (exhibited usually by TF mRNA differential expression or binding in response to the treatment). The second field, Evidence, summarizes the evidence we found for the association. The next field, LLR, contains the pGENMi eQTL+M LLR for the association. The following High Scoring Analyses field contains the experiments that yielded a significant LLR; for eQTL and eQTM, the LLR threshold was 1.74. Since we only examine significant eQTL+M pairs in this analysis, every association should at least have eQTL+M in this field. The last field is Observation which is a paragraph describing literature evidence we found corroborating the (TF, Drug) association or at least providing some evidence as to its validity.

As previously said in the pGENMi Chapter, for each drug, we attempted to validate in literature (with our own results) the top 7 associations with an LLR > 3, resulting in 73 (TF, Drug) associations (see Table 3.7). For those with some, indirect, or direct evidence of association, we summarize the evidence alongside citations to each of the referred papers in the following:

6-MP: (1/7)

Association: NANOG with 6-MP

Observation: Though we failed to corroborate this finding in the literature, we have found NANOG to be associated with at least 4 other drugs. Knockdown of NANOG has been shown to increase sensitivity in cancer cells to cytotoxic agents like cisplatin [122] and reduce malignancy potential [132]. Overexpression of NANOG has also been shown to increase resistance to docetaxel [133]. Thus, while we failed to corroborate this association, NANOG appears to be important in cancer and drug resistance more broadly.

Association: RCOR1 with 6-MP
Observation: The nuclear receptor NR4A2 has been shown to be activated by 6-MP in CV1 and HEK293 cells [134]. In microglia cells and astrocytes, NRF4A2 recruited COREST (RCOR1) in clearing RELA via transcriptional repression [135]. Given that NRFA2 is a

target of 6-MP and interacts with RCOR1 in transcriptional repression of RELA, it stands to reason that RCOR1 is associated with 6-MP, albeit indirectly through NRFA2.

Association: TCF7L2 with 6-MP
Observation: TCF7L2 is known to be involved in drug resistance [136]. In investigating the effect of NR4A3 variant (coding SNP rs12686676) in insulin gene regulation and insulin secretion in -cells, researchers found a gene-gene interaction between the NR4A3 allele and a variant of TCF7L2 (SNP rs7903146). The NR4A3 allele is known to increase insulin secretion, while the TCF7L2 allele decreases secretion. Haplotypes with wild type TCF7L2 and variant NRFA3 do not exhibit significant increase or decrease in insulin secretion compared to haplotypes that are wildtype for both genes. However, haplotypes with variant TCF7L2 and variant NR4A3 show much higher secretion of insulin compared to haplotypes with wildtype NR4A3 and variant TCF7L2, indicating that the NR4A3 can restore insulin secretion in comprised systems with variant TCF7L2. It is widely known that the NR4A subgroup can be activated by 6-MP via the AF-1 domain [137]. Given the interaction between TCF7L2 and NR4A3 as well as NR4A3s activation by 6-MP, we conclude that there is some evidence that TCF7L2 is associated with 6-MP, but not enough to warrant indirect or direct validation status. In a separate study examining the effect of azathioprine, a purine anti-metabolite akin to 6-MP, on carcinogenesis in mice found frameshift mutations in TCF7L2 and seven other genes involved in carcinogenesis when at least one copy of the DNA repair protein MSH2 in the mice was absent [138].

6-TG: (0/4)

All four associations uncorroborated

ARA-C: (5/6)

Association: ZNF274 with Ara-C (Stated in pGENMi main text)

Association: RAD21 with Ara-C
Observation: In our study, siRNA knockdown of RAD21 in Jurkat cell lines shifted the cytotoxicity curve of Cytarabine significantly compared to DMSO controls. Additionally, somatic mutations in RAD21 have been identified as a key driver in acute myeloid leukemia (AML) development; treatment of cytarabine and indarubicin has been shown to achieve remission in patients with cohesin mutations in STAG2, SMC3, and RAD21 [139].

Association: TRIM28 with Ara-C
Observation: TRIM28 is an ataxia telangiectasia mutated (ATM) substrate activated by DNA double strand breaks. In exposing multiple myeloma cells with a combination of gemcitabine (a pyrimidine analog akin to cytarabine) and a purine analog, clofarabine, ATM kinase substrates such as histone 2AX, TRIM28, and p53 were phosphorylated [140], indicating that the ATM pathway (including TRIM28) is activated by this combination. Another study demonstrated that depletion of ATM in HeLa and A549 cells sensitized them to gemcitabine therapy further supporting the importance of this pathway in gemcitabine treatment [141].

Association: ZEB1 with Ara-C
Observation: In a study investigating the role of ZEB1 in differential response to chemotherapy in mantle cell lymphoma (MCL), the investigators knocked down ZEB1 by introducing salinomycin to MCL cells; ZEB1 depends on WNT signaling for expression and salnomycin is a WNT blocker. After ZEB1 knockdown, the researchers noticed increased chemosensitivity of MCL cells to the cytotoxic effects of doxorubicin, cytarabine and gemcitabine [142].

Association: TCF12 with Ara-C
Observation: In a work detailing the effects of cytarabine on ectoderm and mesoderm development in human embryonic stem cells (hESCs), the investigators performed a differential gene expression assay on hESCs. They demonstrated that cytarabine upregulates TCF12 expression in this cell type [143].

ARSENIC: (1/1)

Association: EZH2 with Arsenic
Observation: An analysis of the involvement of JNK and STAT3 in AKT-mediated phosphorylation of EZH2 revealed a critical insight into the role of arsenic in bronchial epithelial cells. It was shown that treating bronchial epithelial cells (BEAS-2B) with arsenic induces phosphorylation of EZH2. This was demonstrated by first transfecting cells with siRNA targeting JNK1; such cells showed a loss of phosphorylation of STAT3, diminished AKT activity, and thus a loss of phosphorylation on serine 21 of EZH2. The researchers also illustrated that arsenic targets the AKT pathway by inducing miRNA-21, a miRNA regulated by STAT3. Silencing this miRNA in transfected arsenic-induced cells diminished phosphorylation of EZH2, while ectopic overexpression of miRNA-21 resulted in EZH2 phosphorylation [144]. With

respect to cell cycle arrest generally, EZH2 has also been implicated as a significant player in SWI/SNF, a family of proteins that consume ATP to remodel nucleosomes, deficient cells; this was shown via inhibition of EZH2 in SMARCB1 (a member of the SWI/SNF family) deficient rhabdoid cells, which led to alterations of H3K27 trimethylation and cytotoxicity [145]

CARBOPLATIN: (1/1)

Association: SP1 with Carboplatin
Observation: It has been established via chromatin immunoprecipitation (ChIP) assays that, in tumor copper deficient cells, SP1 regulates SLC31A1 by binding to the SLC31A1 promoter. The SLC31A1 protein is a membrane protein that transports copper into the cell, and, more importantly, platinum drugs such as cisplatin and carboplatin. Low expressing individuals of SLC31A1 are typically resistant to the chemotherapy of platinum drugs. It has been shown that SLC31A1 can be induced in copper deficient environments through upregulation of SP1, thereby leading to greater chemosensitivity to platinum agents. Such experiments have already been performed with cisplatin and preliminary studies are encouraging for carboplatin therapy [146]. Additionally, a separate study showed that inhibition of SP1 binding to the promoter of BIRC5 (survivin), an antiapoptotic protein highly expressed in cancer and linked to drug resistance, decreased expression of BIRC5 [147], implicating SP1 more broadly in chemotherapy.

CISPLATIN: (7/7)

Association: NANOG with Cisplatin (see Direct Associations)

Association: RELA with Cisplatin
Observation: It was shown in HeLa 57A cells that treatment of cisplatin reduced transcriptional expression of RELA, via luciferase reporter assays. Additionally, further investigation found that cisplatin treatment downregulated BCL2L1. A separate study by the same group demonstrated that RELA is required for activation of BCL2L1 through siRNA knockdown of RELA [148]. Another study examining the interaction of RELA and cisplatin focused on the differential phosphorylation of the T505 residue of RELA by CHEK1 in response to cisplatin treatment in mouse embryonic fibroblasts. The researchers found that T505 phosphorylation induced a proapoptotic form of RELA that could facilitate cell death through transcriptional repression of antiapoptotic target genes such as BCL2L1 [149].

Association: SMARCC1 with Cisplatin
Observation: SMARCC1 is a member of the SWI/SNF family of proteins that remodel nucleosomes. A study of the cytotoxicity of cisplatin in head and neck squamous cell carcinoma revealed that SMARCA4 and SMARCA1 increase sensitivity to cisplatin, via shRNA knockdown experiments [150]. Another study showed that the suppression of SWI/SNF factors (SMARCB1, SMARCC1, SMARCC2, SMARCD1, SMARCD3) via siRNA sensitized cells to cisplatin at the same level as ARID1A and ARID1B suppression in U2OS cells [71]. Furthermore, in yeast, the YJL175W open reading frame deletion that confers cisplatin resistance overlaps with SWI3, the yeast ortholog of SMARCC1 [151].

Association: RUNX3 with Cisplatin
Observation: Examining the role of RUNX3 in gastric cancer chemotherapy, siRNA knockdown of RUNX3 sensitized immortalized stomach mucosal cells (GES-1) and gastric cancer cells (SGC7901) to cisplatin [152]. Another study in hepatocellular carcinoma (HCC) looking at the role of miR-130a/RUNX3/WNT signaling in cisplatin treatment demonstrated through siRNA knockdown of miR-130a and RUNX3 that miR-130a inhibits RUNX3 which activates WNT signaling and results in cellular resistance to cisplatin treatment [153]. Further studies have shown differential expression of RUNX3 in cells treated with cisplatin [154] [155].

Association: WRNIP1 with Cisplatin
Observation: In an investigation of ZRANB3, siRNA knockout in U2OS cells resulted in dramatic sensitization to camptothecin (CPT) and moderate sensitization to cisplatin. The study also elucidated cooperativity between ZRNAB3 and WRNIP1 [156]. In a study of chicken DT40 cells, knockdowns of WRNIP1 did not desensitize the cell to cisplatin, but moderately to CPT; however, negative cells of RAD18, which interacts with WRNIP1, did exhibit high sensitivity to both cisplatin and CPT [157].

Association: NR3C1 with Cisplatin
Observation: In OC-k3 mouse hairs cells, researchers showed that cisplatin treatment upregulated expression of NR3C1 in vitro [158]. However, in a transcriptomic profile of human cellular response to cytotoxic agents, NR3C1 was found to be downregulated by cisplatin treatment [159]. This contradiction may be due to the differing regulatory architectures between the species, as it is likely cisplatin does not interact directly with NR3C1 Additionally, NR3C1 is a glucocorticoid receptor (GR) and glucocorticoids (GC) have been shown to be enhance cytotoxicity [160].

Association: CEBPD with Cisplatin
Observation: In analysis of human urothelial carcinoma cell line NTUB1, researchers examined the following cell sublines resistant to particular drugs to analyze the differences in CEBPD expression: cisplatin (NTUB1/P(14)), gemcitabine (NTUB1/G(1.5)), arsenic trioxide (NTUB1/As(0.5)), and paclitaxel (NTUB1/T(0.017)). CEBPD was only expressed in the cisplatin resistant cell subline, NTUB1/P(14), and in none of the parental cell lines. When treating NTUB1 cells with cisplatin, CEBPD protein and mRNA expression levels were unilaterally elevated across the cell lines. In determining its role in chemoresistance, CEBPD was overexpressed in NTUB1 cells treated with increasing amounts of cisplatin. Compared to controls, CEBPD overexpressed cells were significantly more resistant to cisplatin-induced apoptosis. The paper further articulates that CEBPD reduces cisplatin-induced reactive oxygen species (ROS) production by inducing the expression of Cu/Zn-superoxide dismutase (SOD1) [161]. In a study of ototoxicity in mouse cells, CEBPD expression was also affected by administration of cisplatin [158]. Another study in human found that cisplatin upregulated CEBPD [162], which is consistent with Hour et al.

DOCETAXEL: (1/1)

Association: BATF with Docetaxel (see Direct Associations)

DOXORUBICIN: (3/3)

Association: HNF4G with Doxorubicin (see Direct Associations)

Association: HMGN3 with Doxorubicin
Observations: In the paper by Hanson et al., HMGN3 was silenced via siRNA in the negative breast cancer cell line, MDA-MB-231. The knocked down cell lines were treated with either doxorubicin or epirubicin. Compared to control, HMGN3 knockdowns were more resistant to anthracycline-induced apoptosis in MDA-MB-231 cells, but not BTF459 [44]. In a study profiling gene expression in 30 different cell lines treated with 11 different drugs, HMGN3 expression level was predictive of doxorubicin sensitivity across cells [163].

Association: TCF7L2 with Doxorubicin
Observations: TCF7L2 is a member of the WNT/-Catenin pathway. In a recent paper by Vangipuram et al., the activity of this pathway was investigated in the context of chemore-

sistance in cancer stem-like cells in a neuroblastoma cell line. To start, the investigators segmented the SK-N-SH into resistant (CD133+) and sensitive groups (CD133-) to doxorubicin treatment. Pathway activity scoring showed suppression of the WNT pathway in doxorubicin treated CD133- cells. WNT pathway genes were more differentially expressed in CD133+ cells compared to CD133- cells. When treated with WNT agonists, doxorubicin was very effective in reducing the population of CD133+ cells. This indicates that doxorubicin efficacy is related to the pathway activity of WNT/-Catenin [164]; however, this doesnt necessarily link the drug directly to TCF7L2. A study in the mouse intestinal track demonstrated that doxorubicin affects the expression TCF7L2 [165].

EPIRUBICIN: (3/4)

Association: HNF4G with Epirubicin (see Direct Associations)

Association: HMGN3 with Epirubicin (see HMGN3 with Doxorubicin)

Association: TAL1 with Epirubicin
Observation: TAL1 is expressed frequently in human T-cell acute lymphoblastic leukemia. However, there are aberrations in the expression of TAL1 among leukemia cells. In investigating the cytotoxic effect of TAL1, Bernard et al. transfected human immature T-cell lymphoid cell lines with TAL1. After treating with various cytotoxic agents including doxorubicin, the researchers concluded that the transfectants were more resistant to the cytotoxic effects of these drugs. TAL1 transfectants lacking a DNA binding domain did not show altered sensitivity, revealing that TAL1 binding to DNA is important for this to occur. The researchers, therefore, concluded that TAL1 acts at a late stage of the apoptotic cascade [166]. While doxorubicin is not epirubicin, the two are very similar in structure, belonging to the same drug class (anthracyclines). Additionally, the eQTL+M (TAL1, Doxorubicin) association yielded an LLR score of 2.81, which barely missed our threshold of 3. Therefore, we feel confident generalizing TAL1s association with doxorubicin to epirubicin.

GEMCITABINE: (3/5)

Association: ZNF143 with Gemcitabine
Observations: In looking at respiratory deficient mitochondrial cells, it was shown that cell lines with respiratory dysfunction were more resistant to death via gemcitabine treatment than their normal counterparts. It was also shown that such cells had higher ZNF143 mRNA

levels compared to normal respiratory cells. While gemcitabine treatment with ZNF143 knockdown was not reported, dysfunctional cells showed greater sensitivity to cisplatin after ZNF143 was knocked down [167]. Thus, some evidence exists that ZNF143 can elicit chemoresistance under certain circumstances.

Association: USF1 with Gemcitabine
Observation: Both gemcitabine and ara-C rely on deoxycytidine kinase (DCK) in the first rate limiting steps of the activation of these nucleoside agents in solid tumors and leukemia, respectively [168]. In vivo ChIP assays of HepG2 cells showed the presence of USF1/2 and SP1/2 bound factors to the DCK promoter. Co-transfections in HepG2 showed activation properties of USF1/2 binding and repressive properties of SP1 binding to a DCK-luciferase reporter construct [169]. A separate study showed that DCK expression was predictive of ara-C IC50 in acute myeloid leukemia (AML) cell lines; among the aforementioned regulators of DCK, only USF1 expression was variable and correlated with DCK [168]. Given that USF1 has been shown to regulate a gene important in the pharmacokinetics of both gemcitabine and cytarabine and is predictive of cytarabine IC50 in AML cell lines, we conclude that there is some evidence to suggest USF1 is associated with gemcitabine.

Association: ZNF274 with Gemcitabine
Observation: A gene expression profiling of breast cancer cell lines in response to gemcitabine treatment revealed numerous differentially expressed genes; ZNF274 was upregulated in response to gemcitabine treatment [119].
HYPOXIA: (1/1)

HYPOXIA: (1/1)

Association: PRDM1 with Hypoxia
Observation: PRDM1 is one of two major transcription factors critical for XBP1 expression; it does so by repressing PAX5, itself a repressor of XBP1, thereby enhancing XBP1 expression [170]. As it turns out, XBP1 is essential for hypoxia survival and is required for tumor growth [171]. Though not TF siRNA evidence, we consider regulation of a gene essential for treatment survival to be direct evidence of PRDM1s association with Hypoxia. In multiple myeloma (MM) cells, researchers found that hypoxia induces the downregulation of plasma specific TFs and upregulated stem-cell associated TFs. Among those TFs downregulated in hypoxic MM cells compared to normoxic MM cells was PRMD1 [172]. Paradoxically, transcriptomic profiling in other cells in hypoxic and normoxic conditions revealed upregulation

of PRDM1 mRNA in response to oxygen deprivation [159] [173]. This suggests PRDM1 is affected differently depending on the pathways induced by hypoxia.

NAPQI: (4/7)

Association: MEF2C with NAPQI (see Direct Associations)

Association: NFIC with NAPQI (see Direct Associations)

Association: CCNT2 with NAPQI (see Direct Associations)

Association: UBTF with NAPQI (see Direct Associations)

Association: BCLAF1 with NAPQI (see Direct Associations)
Observation: In a gene expression analysis of precision cut human liver slices, treatment of APAP-induced upregulation of BCLAF1 [105]. BCLAF1 has also been shown to promote cell death generally [174].

OXALIPLATIN: (5/7)

Association: STAT3 with Oxaliplatin (see Direct Associations)

Association: NANOG with Oxaliplatin (see Direct Associations)

Association: TRIM28 with Oxaliplatin
Observation: While there exists little direct evidence linking TRIM28 to oxaliplatin, studies have shown that cisplatin increases phosphorylation [175] [176] and mRNA production (when used in concert with piroxicam) [177] of TRIM28. One prominent study on three non-small cell lung cancer (NSLC) cell lines transformed into tumor-initiating cells (TICs) via stem cell media found impaired phosphorylation of TRIM28 due to irradiation and cisplatin treatment; the researchers hypothesized that the inhibition of TRIM28 phosphorylation might provide a mechanistic explanation for the observed reduction in DNA damage-induced cell cycle arrest and apoptosis in NSCLC TICs [178]. This speaks more broadly to the phosphorylation of TRIM28 desensitizing DNA damaged cells to death rather than to its association with a specific drug. Another very relevant study elucidating the connection between ataxia telangiectasia and Rad3-related (ATR) protein inhibitors and cisplatin demonstrated in-

creased phosphorylation due to mediation by ATR; the same cells sensitized by the ATR inhibitor combined with cisplatin were also sensitized by the ATR inhibitor and oxaliplatin, though to a lesser degree [179]. However, S-phase cell arrest via ATR appeared to be sensitive to cisplatin, and not oxaliplatin [180], limiting the generalizability of the association of discussion to oxaliplatin. Mechanistically, though, it is still plausible for there to be an association. TRIM28 promotes re-sectioning of DNA double strand breaks not protected by -H2AX [181] in murine G1-phase lymphocytes and oxaliplatin has been shown to induce -H2AX [182]. In terms of chemotherapeutic drugs more generally, a study found that SKOV3 cells overexpressed with TRIM28 showed increased resistance to cisplatin and paclitaxel [183], despite the notable differences in mechanism of action between platinum agents and taxane therapies. Thus, while there is sufficient evidence between a TRIM28 and cisplatin association, there are mechanistic reasons to believe it can be generalized to oxaliplatin, a drug of the same family as cisplatin. As for why TRIM28 was not reported as an association with cisplatin, the best score (2.44) didnt make the cutoff threshold of 3 in the eQTL+M analysis.

Association: FOSL1 with Oxaliplatin
Observation: An experiment involving the co-treatment of oxaliplatin with topotecan in bone marrow from rats showed evidence of upregulation of FOSL1 when treated with this cocktail [184]. In addition, FOS signaling has been shown to be activated by oxaliplatin treatment in a variety of cancers [185]. In our analysis, FOS association with oxaliplatin was also noteworthy with a score of 3.61, but it was not in the top 7 associated TFs with oxaliplatin.

Association: ZEB1 with Oxaliplatin
Observation: In colon cancer cell line THC8307/L-OHP, oxaliplatin treatment downregulated ZEB1 expression [186]. ZEB1 has also been shown to increase tumorigenicity by repressing stemness-inhibiting miRNAs like miR-203 [187], which has been shown to increase oxaliplatin resistance in colorectal cancer cells [188].

Association: ATF1 with Oxaliplatin

Association: PPARGC1A with Oxaliplatin
Observation: PPARGC1A siRNA knockdowns were performed in colon cancer liver metastases treated with the chemotherapeutic agents oxaliplatin and 5-fluorouracil (5-FU); knockdown of PPARGC1A prevented chemotherapy-induced oxidative phosphorylation (OXPHOS).

The study concluded that colorectal tumors shift energy production from glycolysis to OXPHOS via the SIRT1/PPARGC1A pathway [189].

PACLITAXEL: (2/3)

Association: GATA1 with Paclitaxel
Observation: Chronic myelogenous leukemia (CML) resulting from a t(9,22) translocation is resistant to paclitaxel treatment. Researchers found that this produced an oncoprotein that activated a GATA1 response element in the promoter of a heat shock protein, HSP70. The siRNA knockdown of HSP7 0sensitized the cell to paclitaxel [190]. Additionally, a murine study of the timeline of paclitaxel-induced cellular changes exhibited upregulation of GATA1 expression [191].

Association: E2F1 with Paclitaxel
Observation: Low dose application of paclitaxel human retinoblastoma cells exhibited upregulation of E2F1 [192]. Additionally, E2F1 overexpression in human osteosarcoma U2OS cells sensitized the cells to paclitaxel [193].

RADIATION: (4/4)

Association: EZH2 with Radiation (see Direct Associations)

Association: ESR1 with Radiation
Observation: In a study examining the molecular mechanisms underlying the transformation of immortalized cells into tumorigenic cells, researchers found that ESR1, while low expressed, was differentially expressed between immortalized mammary epithelial cells and those induced into tumorigenesis via heavy-ion radiation [194]. A study examining the bystander effect, where cells respond to their neighbors, of irradiation in variable estrogen receptor (ER) environments provides clearer proof of this association. The researchers irradiated MDA-MB-231 cells, which are ER negative, and MCF-7 cells, which are ER positive; additionally, they treated both cells with 17-estradiol (E2), which activates ESR1, and tamoxifen, which is an E2 antagonist. MCF-7 cells, which have ESR1, exhibited increased radiosensitivity and bystander response when treated with E2; the effect was diminished by tamoxifen. E2 also increased MCF-7 reactive oxygen species (ROS), absent radiation; however, in MDA-MB-231, neither the bystander response nor ROS increase was observed [195]. Given that ESR1 activation sensitized MCF-7 cells to radiation, we consider this

direct validation.

Association: PML with Radiation
Observation: Mouse embryo fibroblasts (MEFs) proficient and deficient in PML were irradiated to determine the effect of PML in radiation-induced apoptosis. PML deficient cells were much more resistant to radiation, indicating that PML mediates the apoptosis of radiation therapy. Overexpression of PML upon cellular irradiation potentiated c-Jun transcriptional activation and co-activation of c-Jun by PML was observed exclusively in irradiated cells. ChIP experiments of irradiated and unaffected cells showed that binding of c-Jun to its promoter was observed in irradiated, but not unaffected cells. Super shift analysis also showed that the DNA binding ability of c-Jun/ATF-2 was comprised in PML deficient irradiated cells [196]. A separate examination of the interaction between PML and TOPBP1 demonstrated that both co-localize in the nucleus of the cervical cell lines, SiHa, to repair DNA damage after irradiation. Additionally, siRNA PML knockouts exhibited a decrease in radiation-induced TOPBP1 expression, suggesting PML is a regulator of TOPBP1. However, overexpression of PML did not increase mRNA levels of TOPBP1, but did increase TOPBP1 protein expression. Furthermore, pulse-chase labeling experiments indicated that PML increases the half-life of TOPBP1 protein, indicating that this regulation occurs at the post-transcriptional level [197]. A separate analysis in HeLa cells treated by radiation and cisplatin showed upregulation of PML protein in response to treatment, although northern blotting did not indicate a gross increase in mRNA levels. Transfection of p53 into HeLa upregulation of PML with respect to control, indicating that PML is regulated by the p53 pathway [198].

Association: HDAC2 with Radiation
Observation: In non-small cell lung cancer (nsCL) BE1 cells, HDAC1 and HDAC2 expression were highly correlated and significantly higher than normal tissues. Prognosis of patients with low expression of these HDACs was noticeably higher compared to high expression patients. Irradiation of BE1 cells downregulated HDAC1 and HDAC2 expression and upregulated AXIN expression. Knockdowns of HDAC1 and HDAC2 via siRNA upregulated AXIN expression as well. Radiation treatment combined with HDAC knockdown in BE1 cells expressing AXIN increased apoptosis of cells compared to radiation treatment alone [199]. In fact, there is a growing body of literature that suggest combination therapies of HDAC inhibitors with radiation in chemotherapy treatment [200].

RAPAMYCIN: (3/3)

Association: GATA1 with Rapamycin (see Direct Associations)

Association: STAT2 with Rapamycin
Observation: An experiment treating lung adenocarcinoma cells (A549) with rapamycin identified STAT1 interacting with mTOR and increased STAT1 nucelar concentration after rapamycin treatment [201]. While STAT1 is not STAT2, the two form heterodimers and are members of the same protein family. Additionally, the pGENMI STAT1 association with Rapamycin score was 2.49, which though below threshold, is very similar to the STAT2 score. We believe that STAT2 is interpreting the same signal in this context and thus the association is valid. This association was also validated in our own experimental validations of the effect of STAT2 knockdown on rapamycin treated MDA-MB-231 cells.

Association: PHF8 with Rapamycin
Observation: A study showed that PHF8 mRNA and protein levels were downregulated in failing human and mice hearts undergoing hypertrophy. Overexpression of PHF8 identified the aKT/mTOR pathway as a target of PHF8. When this pathway was inhibited by treatment of rapamycin, the phenotype lost by PHF8 deficiency was rescued [202]. As in the previous two associations with rapamycin, we corroborated this association in the laboratory.

TRICIRIBINE: (2/2)

Association: REST with Triciribine
Observation: A study of the regulatory dynamics of REST in small cell lung cancer revealed a putative binding site for REST at the 3 end of the AKT2 UTR. Previous studies had shown that siRNA mediated knockdown of REST correlated with increased AKT phosphorylation. In this particular study, siRNA knockdown of REST resulted in upregulation of AKT2. Temporal gene expression profiling also showed high AKT2 expression in small cell lung cancer cell lines only days after high expression of REST diminished [203]. Though we lack direct evidence, given RESTs regulation of AKT2 and triciribines role as an AKT-inhibitor, we are confident in confirming this association at least partially.

Association: NANOG with Triciribine
Observation: While we lack direct evidence of NANOGs association with triciribine, there is plenty of data linking NANOG with AKT. For instance, a study showed that the PI3K/AKT pathways is important in mediating the regulation of NANOG during differentiation of em-

bryonic carcinoma F9 cells [204]. Another study elucidated that AKT-mediated phosphorylation is crucial for repression of NANOG in differentiating murine embryonal carcinoma cells [205]. Taken together, we believe there exists a plausible connection between NANOG and the AKT inhibitor triciribine.

TEMOZOLOMIDE: (3/7)

Association: FOXM1 with Temozolomide (see Direct Associations)

Association: RELA with Temozolomide (see Direct Associations)

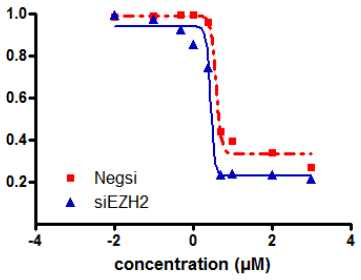
Association: EBF1 with Temozolomide
Observation: In the treatment of glioblastoma, temozolomide works by operating as a methyl donor for the alkylation of the N-7, O-3, and O-6 positions of nucleotide bases, initiating a DNA repair process that cannot undo this level of damage. O-6-methylguanine methyltransferase (MGMT) removes methyl groups from the O-6 position of guanines, thereby rendering temozolomide ineffective [206]. IDH1 mutations have been shown to predict longer survival times via treatment of temozolomide [207]. It has also been shown that EBF1 can bind to both DNA and TET2 and function as a demethylation agent in IDH1 mutants [208]. Thus, EBF1 may have a role to play in differential methylation of the MGMT promoter in IDH1 mutants, and thus have a role to play in cellular response to temozolomide.

Association: ELF1 with Temozolomide
Observation: While we did not find any literature evidence for this association, we validated it through TF siRNA in U251 glioma cells.

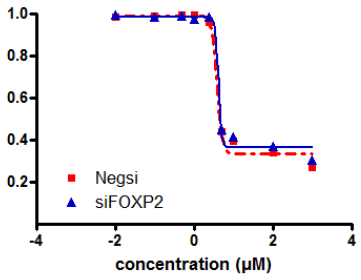
APPENDIX B: PGENMI: FULL CYTOTOXICITY GRAPHS

This appendix contains all 25 pGENMi cytotoxicity graphs, grouped by drug and cell line. Red captioned graphs refer to validated experiments. All experiments were motivated by either a negative control or positive eQTL+M model. For brevity, all experiments are presumed to be positively motivated, unless stated otherwise.

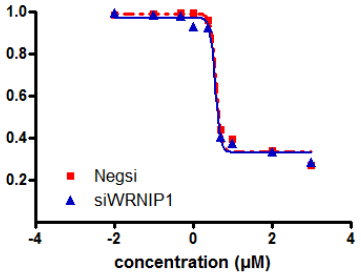
Figure B.1: **(6-MP, Jurkat)**: Cytotoxicity experiments for 5 TFs, in which FOXP2 and EZH2 were negative controls, and EZH2 was validated.



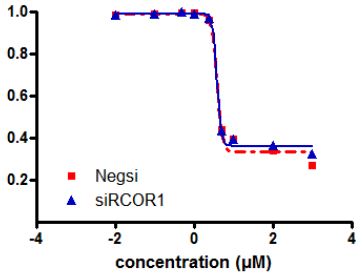
(a) EZH2



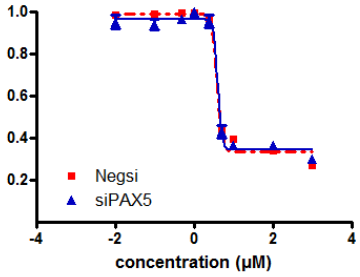
(b) FOXP2



(c) WRNIP1

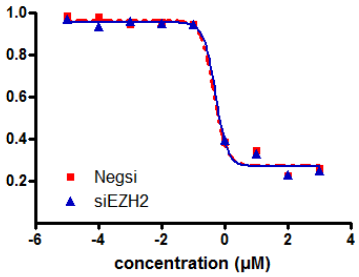


(d) RCOR1

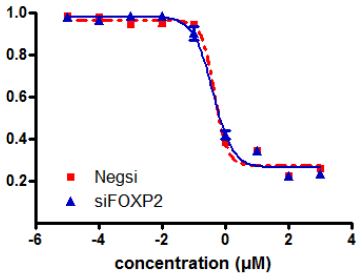


(e) PAX5

Figure B.2: **(6-TG, Jurkat)**: Cytotoxicity experiments for 5 TFs, in which FOXP2 and EZH2 were negative controls and neither validated.

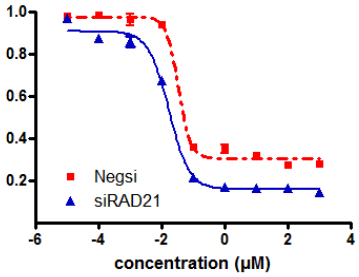


(a) EZH2

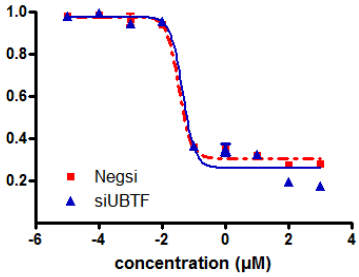


(b) FOXP2

Figure B.3: **(Ara-C, Jurkat)**: Cytotoxicity experiments for UBTF and RAD21 - which validated.

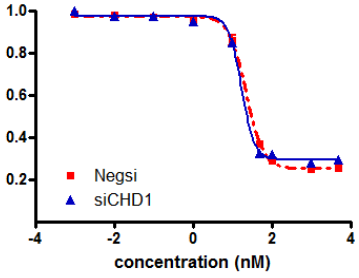


(a) RAD21

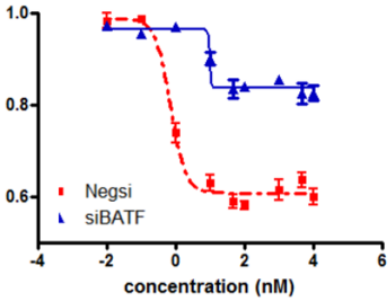


(b) UBTF

Figure B.4: **(Taxanes, MDA-MB-231)**: Cytotoxicity experiments for a negative control, CHD1, and BATF, which validated.

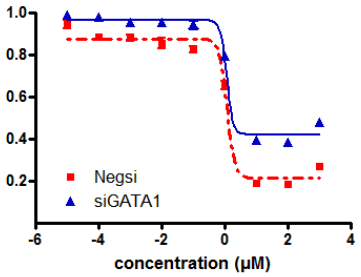


(a) (CHD1, Paclitaxel)

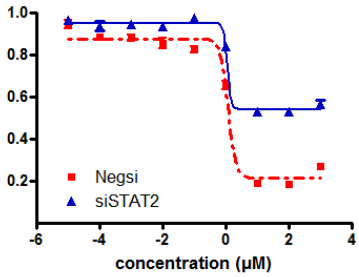


(b) (BATF, Docetaxel)

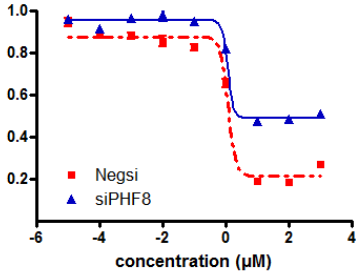
Figure B.5: **(Rapamycin, MDA-MB-231)**: Cytotoxicity experiments for GATA1, STAT2, and PHF8, which all validated.



(a) GATA1

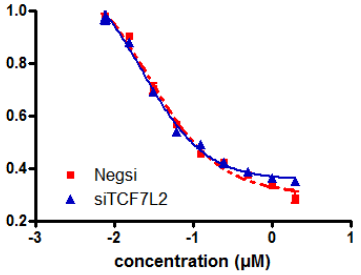


(b) Rapamycin

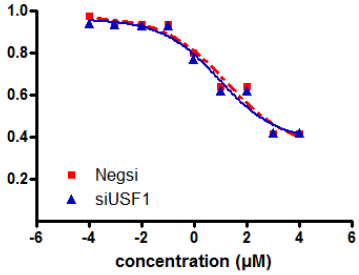


(c) PHF8

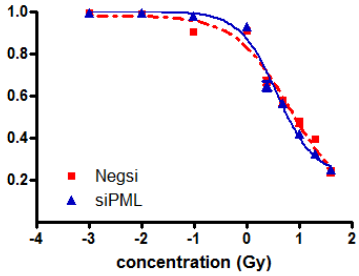
Figure B.6: **(Various, MDA-MB-231)**: Cytotoxicity experiments for (TCF7L2, Epirubicin), (USF1, Gemcitabine), and (PML, Radiation).



(a) (TCF7L2, Epirubicin)

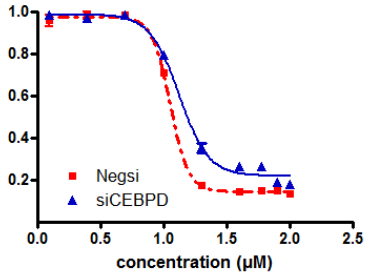


(b) (USF1, Gemcitabine)

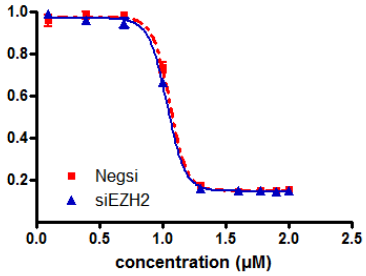


(c) PML

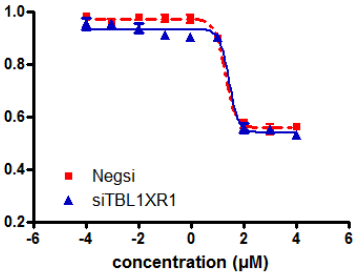
Figure B.7: **(Platinum Agents, MDA-MB-231)**: Cytotoxicity experiments for 3 negative controls, (TBL1XR1, Carboplatin), (EZH2, Cisplatin), (NR3C1, Oxaliplatin), and (CEBPD, Cisplatin), which validated.



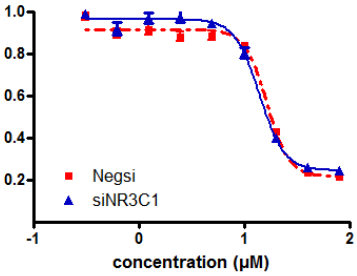
(a) (CEBPD, Cisplatin)



(b) (EZH2, Cisplatin)

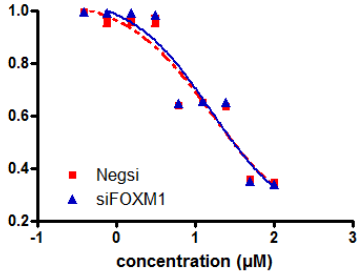


(c) (TBL1XR1,Carboplatin)

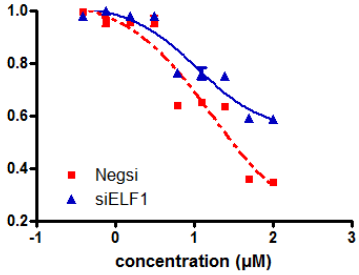


(d) (NR3C1,Oxaliplatin)

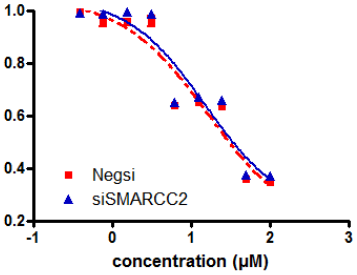
Figure B.8: **(Temozolomide, U251)**: Cytotoxicity experiments for FOXM1, SMARCC2, ZNF263, and ELF1, which validated.



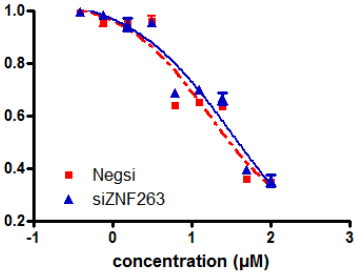
(a) FOXM1



(b) ELF1



(c) SMARCC2



(d) ZNF263

REFERENCES

[1] T. E. P. Consortium, "A user's guide to the encyclopedia of dna elements (encode)," *PLoS Biol*, vol. 9, no. 4, p. e1001046, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21526222

[2] L. Wang and R. M. Weinshilboum, "Pharmacogenomics: candidate gene identification, functional validation and mechanisms," *Human Molecular Genetics*, vol. 17, no. R2, pp. R174–R179, 2008. [Online]. Available: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddn270

[3] L. Wang, H. L. McLeod, and R. M. Weinshilboum, "Genomics and drug response," *New England Journal of Medicine*, vol. 364, no. 12, pp. 1144–1153, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21428770http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3184612http://www.nejm.org/doi/10.1056/NEJMra1010600

[4] E. L. Moen, L. A. Godley, W. Zhang, and M. E. Dolan, "Pharmacogenomics of chemotherapeutic susceptibility and toxicity," *Genome Medicine*, vol. 4, no. 11, pp. 90–90, 2012. [Online]. Available: http://genomemedicine.biomedcentral.com/articles/10.1186/gm391

[5] H. E. Wheeler and M. E. Dolan, "Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation," *Pharmacogenomics*, vol. 13, no. 1, pp. 55–70, 2012. [Online]. Available: https://www.futuremedicine.com/doi/10.2217/pgs.11.121

[6] A. L. Stark, R. J. Hause, L. K. Gorsic, N. N. Antao, S. S. Wong, S. H. Chung, D. F. Gill, H. K. Im, J. L. Myers, K. P. White, R. B. Jones, and M. E. Dolan, "Protein quantitative trait loci identify novel candidates modulating cellular response to chemotherapy," *PLoS Genetics*, vol. 10, no. 4, pp. e1 004 192–e1 004 192, 2014. [Online]. Available: https://dx.plos.org/10.1371/journal.pgen.1004192

[7] K. M. Giacomini, C. M. Brett, R. B. Altman, N. L. Benowitz, M. E. Dolan, D. A. Flockhart, J. A. Johnson, D. F. Hayes, T. Klein, R. M. Krauss, D. L. Kroetz, H. L. McLeod, A. T. Nguyen, M. J. Ratain, M. V. Relling, V. Reus, D. M. Roden, C. A. Schaefer, A. R. Shuldiner, T. Skaar, K. Tantisira, R. F. Tyndale, L. Wang, R. M. Weinshilboum, S. T. Weiss, and I. Zineh, "The pharmacogenetics research network: From snp discovery to clinical drug response," *Clinical Pharmacology & Therapeutics*, vol. 81, no. 3, pp. 328–345, 2007. [Online]. Available: http://doi.wiley.com/10.1038/sj.clpt.6100087

[8] Y. Silberberg, A. Gottlieb, M. Kupiec, E. Ruppin, and R. Sharan, "Large-scale elucidation of drug response pathways in humans," *Journal of Computational Biology*, vol. 19, no. 2, pp. 163–174, 2012. [Online]. Available: http://www.liebertpub.com/doi/10.1089/cmb.2011.0264

[9] R. Huang, A. Wallqvist, N. Thanki, and D. G. Covell, "Linking pathway gene expressions to the growth inhibition response from the national cancer institute's anticancer screen and drug mechanism of action," *The Pharmacogenomics Journal*, vol. 5, no. 6, pp. 381–399, 2005. [Online]. Available: http://www.nature.com/articles/6500331

[10] B. Arjmand, P. Goodarzi, F. Mohamadi-Jahani, K. Falahzadeh, and B. Larijani, "Personalized regenerative medicine," *Acta Medica Iranica*, pp. 144–149, 2017.

[11] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing gwas results: A review of statistical methods and recommendations for their application," *The American Journal of Human Genetics*, vol. 86, no. 1, pp. 6–22, 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0002929709005321

[12] L. Weng, F. Macciardi, A. Subramanian, G. Guffanti, S. G. Potkin, Z. Yu, and X. Xie, "Snp-based pathway enrichment analysis for genome-wide association studies," *BMC Bioinformatics*, vol. 12, no. 1, pp. 99–99, 2011. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-99

[13] M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin, "Powerful snp-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S000292971000248X

[14] J. L. Min, J. M. Taylor, J. B. Richards, T. Watts, F. H. Pettersson, J. Broxholme, K. R. Ahmadi, G. L. Surdulescu, E. Lowy, C. Gieger, C. Newton-Cheh, M. Perola, N. Soranzo, I. Surakka, C. M. Lindgren, J. Ragoussis, A. P. Morris, L. R. Cardon, T. D. Spector, and K. T. Zondervan, "The use of genome-wide eqtl associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits," *PLoS ONE*, vol. 6, no. 7, pp. e22 070–e22 070, 2011. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0022070

[15] L. Li, M. Kabesch, E. Bouzigon, F. Demenais, M. Farrall, M. F. Moffatt, X. Lin, and L. Liang, "Using eqtl weights to improve power for genome-wide association studies: a genetic study of childhood asthma," *Frontiers in Genetics*, vol. 4, pp. 103–103, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23755072http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3668139http://journal.frontiersin.org/article/10.3389/fgene.2013.00103/abstract

[16] E. Choy, R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, M. Rivas, E. T. Dermitzakis, E. Cahir-McFarland, E. Kieff, D. Hafler, M. J. Daly, and D. Altshuler, "Genetic analysis of human traits in vitro: Drug response and gene expression in lymphoblastoid cell lines," *PLoS Genetics*, vol. 4, no. 11, pp. e1 000 287–e1 000 287, 2008. [Online]. Available: http://dx.plos.org/10.1371/journal.pgen.1000287

[17] S. D. Zhao, T. T. Cai, and H. Li, "More powerful genetic association testing via a new statistical framework for integrative genomics," *Biometrics*, vol. 70, no. 4, pp. 881–890, 2014. [Online]. Available: http://doi.wiley.com/10.1111/biom.12206

[18] Y. Ma, Z. Ding, Y. Qian, X. Shi, V. Castranova, E. J. Harner, and L. Guo, "Predicting cancer drug response by proteomic profiling," *Clinical Cancer Research*, vol. 12, no. 15, pp. 4583–4589, 2006. [Online]. Available: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-06-0290

[19] J. Robert, A. Vekris, P. Pourquier, and J. Bonnet, "Predicting drug response based on gene expression," *Critical Reviews in Oncology/Hematology*, vol. 51, no. 3, pp. 205–227, 2004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1040842804001118

[20] J. Zhao, X. S. Zhang, and S. Zhang, "Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 3, no. 2, pp. e102–e102, 2014. [Online]. Available: http://doi.wiley.com/10.1038/psp.2013.79

[21] J. K. Lee, D. M. Havaleshko, H. Cho, J. N. Weinstein, E. P. Kaldjian, J. Karpovich, A. Grimshaw, and D. Theodorescu, "A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery," *Proceedings of the National Academy of Sciences*, vol. 104, no. 32, pp. 13 086–13 091, 2007. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0610292104

[22] R. S. Huang, S. Duan, S. J. Shukla, E. O. Kistner, T. A. Clark, T. X. Chen, A. C. Schweitzer, J. E. Blume, and M. E. Dolan, "Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 427–437, 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S000292970761341X

[23] R. S. Huang, S. Duan, W. K. Bleibel, E. O. Kistner, W. Zhang, T. A. Clark, T. X. Chen, A. C. Schweitzer, J. E. Blume, N. J. Cox, and M. E. Dolan, "A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity," *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9758–9763, 2007. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0703736104

[24] N. Niu, Y. Qin, B. L. Fridley, J. Hou, K. R. Kalari, M. Zhu, T. Y. Wu, G. D. Jenkins, A. Batzler, and L. Wang, "Radiation pharmacogenomics: A genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines," *Genome Research*, vol. 20, no. 11, pp. 1482–1492, 2010. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.107672.110

[25] T. Pastinen, "Genome-wide allele-specific analysis: insights into regulatory variation," *Nature Reviews Genetics*, vol. 11, no. 8, pp. 533–538, 2010. [Online]. Available: http://www.nature.com/articles/nrg2815

[26] E. S. Iversen, G. Lipton, M. A. Clyde, and A. N. A. Monteiro, "Functional annotation signatures of disease susceptibility loci improve snp association analysis," *BMC Genomics*, vol. 15, no. 1, pp. 398–398, 2014. [Online]. Available: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-398

[27] C. D. Brown, L. M. Mangravite, and B. E. Engelhardt, "Integrative modeling of eqtls and cis-regulatory elements suggests mechanisms underlying cell type specificity of eqtls," *PLoS Genetics*, vol. 9, no. 8, pp. e1 003 649–e1 003 649, 2013. [Online]. Available: http://dx.plos.org/10.1371/journal.pgen.1003649

[28] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proc Natl Acad Sci U S A*, vol. 107, no. 33, pp. 14 621–6, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20679242

[29] E. Gregori-Puigjane, V. Setola, J. Hert, B. A. Crews, J. J. Irwin, E. Lounkine, L. Marnett, B. L. Roth, and B. K. Shoichet, "Identifying mechanism-of-action targets for drugs and probes," *Proc Natl Acad Sci U S A*, vol. 109, no. 28, pp. 11 178–83, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22711801

[30] M. Schenone, V. Dancik, B. K. Wagner, and P. A. Clemons, "Target identification and mechanism of action in chemical biology and drug discovery," *Nat Chem Biol*, vol. 9, no. 4, pp. 232–240, 2013. [Online]. Available: http://dx.doi.org/10.1038/nchembio.1199

[31] A. G. Madian, H. E. Wheeler, R. B. Jones, and M. E. Dolan, "Relating human genetic variation to variation in drug responses," *Trends Genet*, vol. 28, no. 10, pp. 487–95, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22840197

[32] M. R. Nelson, T. Johnson, L. Warren, A. R. Hughes, S. L. Chissoe, C.-F. Xu, and D. M. Waterworth, "The genetics of drug efficacy: opportunities and challenges," *Nat Rev Genet*, vol. 17, no. 4, pp. 197–206, 2016. [Online]. Available: http://dx.doi.org/10.1038/nrg.2016.12

[33] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, J. Aspesi, P., M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–7, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22460905

[34] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell, "Cosmic: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D805–11, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25355519

[35] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, "Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D955–61, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23180760

[36] M. G. Rees, B. Seashore-Ludlow, J. H. Cheah, D. J. Adams, E. V. Price, S. Gill, S. Javaid, M. E. Coletti, V. L. Jones, N. E. Bodycombe, C. K. Soule, B. Alexander, A. Li, P. Montgomery, J. D. Kotz, C. S. Hon, B. Munoz, T. Liefeld, V. Dancik, D. A. Haber, C. B. Clish, J. A. Bittker, M. Palmer, B. K. Wagner, P. A. Clemons, A. F. Shamji, and S. L. Schreiber, "Correlating chemical sensitivity and basal gene expression reveals mechanism of action," *Nat Chem Biol*, vol. 12, no. 2, pp. 109–16, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26656090

[37] M. J. Jones, A. P. Fejes, and M. S. Kobor, "Dna methylation, genotype and gene expression: who is driving and who is along for the ride?" *Genome Biol*, vol. 14, no. 7, p. 126, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23899167

[38] F. Fontaine, J. Overman, and M. Francois, "Pharmacological manipulation of transcription factor protein-protein interactions: opportunities and obstacles," *Cell Regen (Lond)*, vol. 4, no. 1, p. 2, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25848531

[39] K. A. Papavassiliou and A. G. Papavassiliou, "Transcription factor drug targets," *Journal of Cellular Biochemistry*, vol. 117, no. 12, pp. 2693–2696, 2016. [Online]. Available: http://dx.doi.org/10.1002/jcb.25605

[40] M. S. Redell and D. J. Tweardy, "Targeting transcription factors in cancer: Challenges and evolving strategies," *Drug Discov Today Technol*, vol. 3, no. 3, pp. 261–7, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24980527

[41] A. M. Redmond and J. S. Carroll, "Defining and targeting transcription factors in cancer," *Genome Biol*, vol. 10, no. 7, p. 311, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19664186

[42] J. E. Yeh, P. A. Toniolo, and D. A. Frank, "Targeting transcription factors: promising new strategies for cancer therapy," *Curr Opin Oncol*, vol. 25, no. 6, pp. 652–8, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24048019

[43] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[44] C. Hanson, J. Cairns, L. Wang, and S. Sinha, "Computational discovery of transcription factors associated with drug response," *The pharmacogenomics journal*, vol. 16, no. 6, pp. 573–582, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26503816https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4848185/

[45] T.-Y. Wu, B. L. Fridley, G. D. Jenkins, A. Batzler, L. Wang, and R. M. Weinshilboum, "Mycophenolic acid response biomarkers: A cell line model system-based genome-wide screen," *International Immunopharmacology*, vol. 11, no. 8, pp. 1057–1064, 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1567576911001196

[46] A. M. Moyer, B. L. Fridley, G. D. Jenkins, A. J. Batzler, L. L. Pelleymounter, K. R. Kalari, Y. Ji, Y. Chai, K. K. S. Nordgren, and R. M. Weinshilboum, "Acetaminophen-napqi hepatotoxicity: A cell line model system genome-wide association study," *Toxicological Sciences*, vol. 120, no. 1, pp. 33–41, 2011. [Online]. Available: https://academic.oup.com/toxsci/article-lookup/doi/10.1093/toxsci/kfq375

[47] B. L. Fridley, A. Batzler, L. Li, F. Li, A. Matimba, G. D. Jenkins, Y. Ji, L. Wang, and R. M. Weinshilboum, "Gene set analysis of purine and pyrimidine antimetabolites cancer therapies," *Pharmacogenetics and Genomics*, vol. 21, no. 11, pp. 701–712, 2011. [Online]. Available: https://insights.ovid.com/crossref?an=01213011-201111000-00003

[48] L. Li, B. L. Fridley, K. Kalari, N. Niu, G. Jenkins, A. Batzler, R. P. Abo, D. Schaid, and L. Wang, "Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines," *BMC genomics*, vol. 15, no. 1, p. 93, 2014.

[49] X. L. Tan, A. M. Moyer, B. L. Fridley, D. J. Schaid, N. Niu, A. J. Batzler, G. D. Jenkins, R. P. Abo, L. Li, J. M. Cunningham, Z. Sun, P. Yang, and L. Wang, "Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy," *Clinical Cancer Research*, vol. 17, no. 17, pp. 5801–5811, 2011. [Online]. Available: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-11-1133

[50] B. L. Fridley, R. Abo, X.-L. Tan, G. D. Jenkins, A. Batzler, A. M. Moyer, J. M. Biernacka, and L. Wang, "Integrative gene set analysis: Application to platinum pharmacogenomics," *OMICS: A Journal of Integrative Biology*, vol. 18, no. 1, pp. 34–41, 2014. [Online]. Available: http://www.liebertpub.com/doi/10.1089/omi.2013.0099

[51] N. Niu, D. J. Schaid, R. P. Abo, K. Kalari, B. L. Fridley, Q. Feng, G. Jenkins, A. Batzler, A. G. Brisbin, J. M. Cunningham, L. Li, Z. Sun, P. Yang, and L. Wang, "Genetic association with overall survival of taxane-treated lung cancer patients - a genome-wide association study in human lymphoblastoid cell lines followed by a clinical association study," *BMC Cancer*, vol. 12, no. 1, pp. 422–422, 2012. [Online]. Available: http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-12-422

[52] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. R. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, M. Kellis, D. Acevedo, R. Auburn, G. Barber, H. J. Bellen et al., "Identification of functional elements and regulatory circuits by drosophila modencode," *Science*, vol. 330, no. 6012, pp. 1787–1797, 2010. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1198374

[53] L. Li, B. Fridley, K. Kalari, G. Jenkins, A. Batzler, S. Safgren, M. Hildebrandt, M. Ames, D. Schaid, and L. Wang, "Gemcitabine and cytosine arabinoside cytotoxicity: Association with lymphoblastoid cell expression," *Cancer Research*, vol. 68, no. 17, pp. 7050–7058, 2008. [Online]. Available: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-08-0405

[54] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102

[55] I. Sur, S. Tuupanen, T. Whitington, L. A. Aaltonen, and J. Taipale, "Lessons from functional analysis of genome-wide association studies," *Cancer Research*, vol. 73, no. 14, pp. 4180–4184, 2013. [Online]. Available: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-13-0789

[56] B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, and E. T. Dermitzakis, "Patterns of cis regulatory variation in diverse human populations," *PLoS Genetics*, vol. 8, no. 4, pp. e1 002 639–e1 002 639, 2012. [Online]. Available: https://dx.plos.org/10.1371/journal.pgen.1002639

[57] R. Ohlsson, R. Renkawitz, and V. Lobanenkov, "Ctcf is a uniquely versatile transcription regulator linked to epigenetics and disease," *Trends in Genetics*, vol. 17, no. 9, pp. 520–527, 2001. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168952501023666

[58] D.-Y. Kim, M.-J. Kim, H.-B. Kim, J.-W. Lee, J.-H. Bae, D.-W. Kim, C.-D. Kang, and S.-H. Kim, "Suppression of multidrug resistance by treatment with trail in human ovarian and breast cancer cells with high level of c-myc," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1812, no. 7, pp. 796–805, 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925443911000846

[59] M. Bonovich, M. Olive, E. Reed, B. O'Connell, and C. Vinson, "Adenoviral delivery of a-fos, an ap-1 dominant negative, selectively inhibits drug resistance in two human cancer cell lines," *Cancer Gene Therapy*, vol. 9, no. 1, pp. 62–70, 2002. [Online]. Available: http://www.nature.com/articles/7700409

[60] X. Li, R. Yao, L. Yue, W. Qiu, W. Qi, S. Liu, Y. Yao, and J. Liang, "Foxm1 mediates resistance to docetaxel in gastric cancer ¡i¿via¡/i¿ up-regulating stathmin," *Journal of Cellular and Molecular Medicine*, vol. 18, no. 5, pp. 811–823, 2014. [Online]. Available: http://doi.wiley.com/10.1111/jcmm.12216

[61] K. Okada, Y. Fujiwara, T. Takahashi, Y. Nakamura, S. Takiguchi, K. Nakajima, H. Miyata, M. Yamasaki, Y. Kurokawa, M. Mori, and Y. Doki, "Overexpression of forkhead box m1 transcription factor (foxm1) is a potential prognostic marker and enhances chemoresistance for docetaxel in gastric cancer," *Annals of Surgical Oncology*, vol. 20, no. 3, pp. 1035–1043, 2013. [Online]. Available: http://link.springer.com/10.1245/s10434-012-2680-0

[62] Q. Zhao, J. Zhang, D. T. Hendricks, and X. Zhao, "Gro and its downstream effector egr1 regulate cisplatin-induced apoptosis in whco1 cells," *Oncology Reports*, vol. 25, no. 4, pp. 1031–7, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21271225http://www.spandidos-publications.com/or/25/4/1031

[63] J. O. Park, C. A. Lopez, V. K. Gupta, C. K. Brown, H. J. Mauceri, T. E. Darga, A. Manan, S. Hellman, M. C. Posner, D. W. Kufe, and R. R. Weichselbaum, "Transcriptional control of viral gene therapy by cisplatin," *Journal of Clinical Investigation*, vol. 110, no. 3, pp. 403–410, 2002. [Online]. Available: http://www.jci.org/articles/view/15548

[64] W.-d. Wang, R. Li, Z.-t. Chen, D.-z. Li, Y.-z. Duan, and Z.-h. Cao, "Cisplatin-controlled p53 gene therapy for human non-small cell lung cancer xenografts in athymic nude mice via the carg elements," *Cancer Science*, vol. 96, no. 10, pp. 706–712, 2005. [Online]. Available: http://doi.wiley.com/10.1111/j.1349-7006.2005.00105.x

[65] D. Roberts, J. Schick, S. Conway, S. Biade, P. B. Laub, J. P. Stevenson, T. C. Hamilton, P. J. O'Dwyer, and S. W. Johnson, "Identification of genes associated with platinum drug sensitivity and resistance in human ovarian cancer cells," *British Journal of Cancer*, vol. 92, no. 6, pp. 1149–1158, 2005. [Online]. Available: http://www.nature.com/articles/6602447

[66] T. Kaur, D. Mukherjea, K. Sheehan, S. Jajoo, L. P. Rybak, and V. Ramkumar, "Short interfering rna against stat1 attenuates cisplatin-induced ototoxicity in the rat by suppressing inflammation," *Cell Death & Disease*, vol. 2, no. 7, pp. e180–e180, 2011. [Online]. Available: http://www.nature.com/articles/cddis201163

[67] N. C. Schmitt, E. W. Rubel, and N. M. Nathanson, "Cisplatin-induced hair cell death requires stat1 and is attenuated by epigallocatechin gallate," *Journal of Neuroscience*, vol. 29, no. 12, pp. 3843–3851, 2009. [Online]. Available: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5842-08.2009

[68] X. Li, H. Wang, X. Lu, and B. Di, "Stat3 blockade with shrna enhances radiosensitivity in hep-2 human laryngeal squamous carcinoma cells," *Oncology reports*, vol. 23, no. 2, pp. 345–53, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20043094

[69] L. Gao, F.-S. Li, X.-H. Chen, Q.-W. Liu, J.-B. Feng, Q.-J. Liu, and X. Su, "Radiation induces phosphorylation of stat3 in a dose- and time-dependent manner," *Asian Pacific Journal of Cancer Prevention*, vol. 15, no. 15, pp. 6161–6164, 2014. [Online]. Available: http://koreascience.or.kr/journal/view.jsp?kj=POCPA9&py=2014&vnc=v15n15&sp=6161

[70] J. Halliday, K. Helmy, S. S. Pattwell, K. L. Pitter, Q. LaPlant, T. Ozawa, and E. C. Holland, "In vivo radiation response of proneural glioma characterized by protective p53 transcriptional program and proneural-mesenchymal shift," *Proceedings of the National Academy of Sciences*, vol. 111, no. 14, pp. 5248–5253, 2014. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1321014111

[71] R. Watanabe, A. Ui, S. i. Kanno, H. Ogiwara, T. Nagase, T. Kohno, and A. Yasui, "Swi/snf factors required for cellular resistance to dna damage include arid1a and arid1b and show interdependent protein stability," *Cancer Research*, vol. 74, no. 9, pp. 2465–2475, 2014. [Online]. Available: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-13-3608

[72] D. Kashatus, "Expression of the bcl-3 proto-oncogene suppresses p53 activation," *Genes & Development*, vol. 20, no. 2, pp. 225–235, 2006. [Online]. Available: http://www.genesdev.org/cgi/doi/10.1101/gad.1352206

[73] E. Paolicchi, F. Crea, W. L. Farrar, J. E. Green, and R. Danesi, "Histone lysine demethylases in breast cancer," *Crit Rev Oncol Hematol*, vol. 86, no. 2, pp. 97–103, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23266085

[74] B. J. Klein, L. Piao, Y. Xi, H. Rincon-Arano, S. B. Rothbart, D. Peng, H. Wen, C. Larson, X. Zhang, X. Zheng, M. A. Cortazar, P. V. Pena, A. Mangan, D. L. Bentley, B. D. Strahl, M. Groudine, W. Li, X. Shi, and T. G. Kutateladze, "The histone-h3k4-specific demethylase kdm5b binds to its substrate and product through distinct phd fingers," *Cell Rep*, vol. 6, no. 2, pp. 325–35, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24412361

[75] H. J. Noh, K. A. Kim, and K. C. Kim, "p53 down-regulates setdb1 gene expression during paclitaxel induced-cell death," *Biochem Biophys Res Commun*, vol. 446, no. 1, pp. 43–8, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24565839

[76] R. Hill, M. Rabb, P. A. Madureira, D. Clements, S. A. Gujar, D. M. Waisman, C. A. Giacomantonio, and P. W. Lee, "Gemcitabine-mediated tumour regression and p53-dependent gene expression: implications for colon and pancreatic cancer therapy," *Cell Death Dis*, vol. 4, p. e791, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24008735

[77] F. Muramatsu, H. Kidoya, H. Naito, S. Sakimoto, and N. Takakura, "microrna-125b inhibits tube formation of blood vessels through translational suppression of ve-cadherin," *Oncogene*, vol. 32, no. 4, pp. 414–21, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22391569

[78] M. P. Puissegur, R. Eichner, C. Quelen, E. Coyaud, B. Mari, K. Lebrigand, C. Broccardo, F. Nguyen-Khac, M. Bousquet, and P. Brousset, "B-cell regulator of immunoglobulin heavy-chain transcription (bright)/arid3a is a direct target of the oncomir microrna-125b in progenitor b-cells," *Leukemia*, vol. 26, no. 10, pp. 2224–32, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22469780

[79] L. Xu, S. Beckebaum, S. Iacob, G. Wu, G. M. Kaiser, A. Radtke, C. Liu, I. Kabar, H. H. Schmidt, X. Zhang, M. Lu, and V. R. Cicinnati, "Microrna-101 inhibits human hepatocellular carcinoma progression through ezh2 downregulation and increased cytostatic drug sensitivity," *J Hepatol*, vol. 60, no. 3, pp. 590–8, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24211739

[80] K. A. Harradine, M. Kassner, D. Chow, M. Aziz, D. D. Von Hoff, J. B. Baker, H. Yin, and R. J. Pelham, "Functional genomics reveals diverse cellular processes that modulate tumor cell response to oxaliplatin," *Mol Cancer Res*, vol. 9, no. 2, pp. 173–82, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21169384

[81] W. M. Abdel-Rahman, S. Knuutila, P. Peltomaki, D. J. Harrison, and S. A. Bader, "Truncation of mbd4 predisposes to reciprocal chromosomal translocations and alters the response to therapeutic agents in colon cancer cells," *DNA Repair (Amst)*, vol. 7, no. 2, pp. 321–8, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18162445

[82] I. Nordentoft, L. Dyrskjot, J. S. Bodker, P. J. Wild, A. Hartmann, S. Bertz, J. Lehmann, T. F. Orntoft, and K. Birkenkamp-Demtroder, "Increased expression of transcription factor tfap2alpha correlates with chemosensitivity in advanced bladder cancer," *BMC Cancer*, vol. 11, p. 135, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21489314

[83] L. A. McPherson, V. R. Baichwal, and R. J. Weigel, "Identification of erf-1 as a member of the ap2 transcription factor family," *Proc Natl Acad Sci U S A*, vol. 94, no. 9, pp. 4342–7, 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9113991

[84] J. Domingo-Domenech, C. Oliva, A. Rovira, J. Codony-Servat, M. Bosch, X. Filella, C. Montagut, M. Tapia, C. Campas, L. Dang, M. Rolfe, J. S. Ross, P. Gascon, J. Albanell, and B. Mellado, "Interleukin 6, a nuclear factor-kappab target, predicts resistance to docetaxel in hormone-independent prostate cancer and nuclear factor-kappab inhibition by ps-1145 enhances docetaxel antitumor activity," *Clin Cancer Res*, vol. 12, no. 18, pp. 5578–86, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17000695

[85] Y. Inoue, M. Gika, T. Abiko, T. Oyama, Y. Saitoh, H. Yamazaki, M. Nakamura, Y. Abe, M. Kawamura, and K. Kobayashi, "Bcl-2 overexpression enhances in vitro sensitivity against docetaxel in non-small cell lung cancer," *Oncol Rep*, vol. 13, no. 2, pp. 259–64, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15643508

[86] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–9, 2006. [Online]. Available: http://dx.doi.org/10.1038/ng1847

[87] J. J Joo, J. Sul, B. Han, C. Ye, and E. Eskin, "Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies," *Genome Biology*, vol. 15, no. 4, pp. r61–r61, 2014. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-4-r61

[88] D. Chung, C. Yang, C. Li, J. Gelernter, and H. Zhao, "Gpa: A statistical approach to prioritizing gwas results by integrating pleiotropy and annotation," *PLoS Genetics*, vol. 10, no. 11, pp. e1 004 787–e1 004 787, 2014. [Online]. Available: https://dx.plos.org/10.1371/journal.pgen.1004787

[89] P. Sudarsanam and B. A. Cohen, "Single nucleotide variants in transcription factors associate more tightly with phenotype than with gene expression," *PLoS Genetics*, vol. 10, no. 5, pp. e1 004 325–e1 004 325, 2014. [Online]. Available: https://dx.plos.org/10.1371/journal.pgen.1004325

[90] R. Hause, A. Stark, N. Antao, L. Gorsic, S. Chung, C. Brown, S. Wong, D. Gill, J. Myers, L. To, K. White, M. . Dolan, and R. Jones, "Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels," *The American Journal of Human Genetics*, vol. 95, no. 2, pp. 194–208, 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0002929714003140

[91] D. J. Gaffney, J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "Dissecting the regulatory architecture of gene expression qtls," *Genome Biology*, vol. 13, no. 1, pp. R7–R7, 2012. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-1-r7

[92] A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson, and D. Koller, "Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals," *Genome Research*, vol. 24, no. 1, pp. 14–24, 2014. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.155192.113

[93] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, 2009. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1181369

[94] D. Xie, A. Boyle, L. Wu, J. Zhai, T. Kawli, and M. Snyder, "Dynamic trans-acting factor colocalization in human cells," *Cell*, vol. 155, no. 3, pp. 713–724, 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867413012178

[95] C. Hanson, J. Cairns, L. Wang, and S. Sinha, "Principled multi-omic analysis reveals gene regulatory mechanisms of phenotype variation," *Genome Res*, vol. 28, no. 8, pp. 1207–1216, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29898900

[96] H. Heyn, S. Moran, I. Hernando-Herraez, S. Sayols, A. Gomez, J. Sandoval, D. Monk, K. Hata, T. Marques-Bonet, L. Wang, and M. Esteller, "Dna methylation contributes to natural human variation," *Genome Res*, vol. 23, no. 9, pp. 1363–72, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23908385

[97] R. Cloney, "Complex traits: Integrating gene variation and expression to understand complex traits," *Nat Rev Genet*, vol. 17, no. 4, p. 194, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26900024

[98] B. L. Browning and Z. Yu, "Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies," *Am J Hum Genet*, vol. 85, no. 6, pp. 847–61, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19931040

[99] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation plink: rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, p. 7, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25722852

[100] N. Zhang, X. Wu, L. Yang, F. Xiao, H. Zhang, A. Zhou, Z. Huang, and S. Huang, "Foxm1 inhibition sensitizes resistant glioblastoma cells to temozolomide by downregulating the expression of dna-repair gene rad51," *Clin Cancer Res*, vol. 18, no. 21, pp. 5961–71, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22977194

[101] M. M. Shahzad, L. S. Mangala, H. D. Han, C. Lu, J. Bottsford-Miller, M. Nishimura, E. M. Mora, J. W. Lee, R. L. Stone, C. V. Pecot, D. Thanapprapasr, J. W. Roh, P. Gaur, M. P. Nair, Y. Y. Park, N. Sabnis, M. T. Deavers, J. S. Lee, L. M. Ellis, G. Lopez-Berestein, W. J. McConathy, L. Prokai, A. G. Lacko, and A. K. Sood, "Targeted delivery of small interfering rna using reconstituted high-density lipoprotein nanoparticles," *Neoplasia*, vol. 13, no. 4, pp. 309–19, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21472135

[102] A. Bavelloni, I. Faenza, M. Aluigi, A. Ferri, A. Toker, N. M. Maraldi, and S. Marmiroli, "Inhibition of phosphoinositide 3-kinase impairs pre-commitment cell cycle traverse and prevents differentiation in erythroleukaemia cells," *Cell Death Differ*, vol. 7, no. 1, pp. 112–7, 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10713726

[103] C. B. Kobarg, J. Kobarg, D. P. Crosara-Alberto, T. H. Theizen, and K. G. Franchini, "Mef2c dna-binding activity is inhibited through its interaction with the regulatory protein ki-1/57," *FEBS Lett*, vol. 579, no. 12, pp. 2615–22, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15862299

[104] R. P. Beyer, R. C. Fry, M. R. Lasarev, L. A. McConnachie, L. B. Meira, V. S. Palmer, C. L. Powell, P. K. Ross, T. K. Bammler, B. U. Bradford, A. B. Cranson, M. L. Cunningham, R. D. Fannin, G. M. Higgins, P. Hurban, R. J. Kayton, K. F. Kerr, O. Kosyk, E. K. Lobenhofer, S. O. Sieber, P. A. Vliet, B. K. Weis, R. Wolfinger, C. G. Woods, J. H. Freedman, E. Linney, W. K. Kaufmann, T. J. Kavanagh, R. S. Paules, I. Rusyn, L. D. Samson, P. S. Spencer, W. Suk, R. J. Tennant, H. Zarbl, and C. Members of the Toxicogenomics Research, "Multicenter study of acetaminophen hepatotoxicity reveals the importance of biological endpoints in genomic analyses," *Toxicol Sci*, vol. 99, no. 1, pp. 326–37, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17562736

[105] M. G. Elferink, P. Olinga, E. M. van Leeuwen, S. Bauerschmidt, J. Polman, W. G. Schoonen, S. H. Heisterkamp, and G. M. Groothuis, "Gene expression analysis of precision-cut human liver slices indicates stable expression of adme-tox related genes," *Toxicol Appl Pharmacol*, vol. 253, no. 1, pp. 57–69, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21420995

[106] J. S. Moffit, P. H. Koza-Taylor, R. D. Holland, M. S. Thibodeau, R. D. Beger, M. P. Lawton, and J. E. Manautou, "Differential gene expression in mouse liver associated with the hepatoprotective effect of clofibrate," *Toxicol Appl Pharmacol*, vol. 222, no. 2, pp. 169–79, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17585979

[107] J. M. Prot, A. Bunescu, B. Elena-Herrmann, C. Aninat, L. C. Snouber, L. Griscom, F. Razan, F. Y. Bois, C. Legallais, C. Brochot, A. Corlu, M. E. Dumas, and E. Leclerc, "Predictive toxicology using systemic biology and liver microfluidic "on chip" approaches: application to acetaminophen injury," *Toxicol Appl Pharmacol*, vol. 259, no. 3, pp. 270–80, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22230336

[108] J. Tel, S. V. Hato, R. Torensma, S. I. Buschow, C. G. Figdor, W. J. Lesterhuis, and I. J. de Vries, "The chemotherapeutic drug oxaliplatin differentially affects blood dc function dependent on environmental cues," *Cancer Immunol Immunother*, vol. 61, no. 7, pp. 1101–11, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22193989

[109] S. Cross-Knorr, S. Lu, K. Perez, S. Guevara, K. Brilliant, C. Pisano, P. J. Quesenberry, M. B. Resnick, and D. Chatterjee, "Rkip phosphorylation and stat3 activation is inhibited by oxaliplatin and camptothecin and are associated with poor prognosis in stage ii colon cancer patients," *BMC Cancer*, vol. 13, p. 463, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24098947

[110] W. J. Sheng, H. Jiang, D. L. Wu, and J. H. Zheng, "Early responses of the stat3 pathway to platinum drugs are associated with cisplatin resistance in epithelial ovarian cancer," *Braz J Med Biol Res*, vol. 46, no. 8, pp. 650–8, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23969971

[111] S. Caporali, L. Levati, G. Graziani, A. Muzi, M. G. Atzori, E. Bonmassar, G. Palmieri, P. A. Ascierto, and S. D'Atri, "Nf-kappab is activated in response to temozolomide in an akt-dependent manner and confers protection against the growth suppressive effect of the drug," *J Transl Med*, vol. 10, p. 252, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23259744

[112] B. Yamini, X. Yu, M. E. Dolan, M. H. Wu, T. E. Darga, D. W. Kufe, and R. R. Weichselbaum, "Inhibition of nuclear factor-kappab activity by temozolomide involves o6-methylguanine induced inhibition of p65 dna binding," *Cancer Res*, vol. 67, no. 14, pp. 6889–98, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17638900

[113] Y. Y. Zaytseva, J. D. Valentino, P. Gulhati, and B. M. Evers, "mtor inhibitors in cancer therapy," *Cancer Lett*, vol. 319, no. 1, pp. 1–7, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22261336

[114] Z. J. Liu, J. Italiano, J., F. Ferrer-Marin, R. Gutti, M. Bailey, B. Poterjoy, L. Rimsza, and M. Sola-Visner, "Developmental differences in megakaryocytopoiesis are associated with up-regulated tpo signaling through mtor and elevated gata-1 levels in neonatal megakaryocytes," *Blood*, vol. 117, no. 15, pp. 4106–17, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21304100

[115] L. O. Andrieux, A. Fautrel, A. Bessard, A. Guillouzo, G. Baffet, and S. Langouet, "Gata-1 is essential in egf-mediated induction of nucleotide excision repair activity and ercc1 expression through erk2 in human hepatoma cells," *Cancer Res*, vol. 67, no. 5, pp. 2114–23, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17332341

[116] M. I. James, C. Iwuji, G. Irving, A. Karmokar, J. A. Higgins, N. Griffin-Teal, A. Thomas, P. Greaves, H. Cai, S. R. Patel, B. Morgan, A. Dennison, M. Metcalfe, G. Garcea, D. M. Lloyd, D. P. Berry, W. P. Steward, L. M. Howells, and K. Brown, "Curcumin inhibits cancer stem cell phenotypes in ex vivo models of colorectal liver metastases, and is clinically safe and tolerable in combination with folfox chemotherapy," *Cancer Lett*, vol. 364, no. 2, pp. 135–41, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25979230

[117] S. Y. Ju, C. Y. Huang, W. C. Huang, and Y. Su, "Identification of thiostrepton as a novel therapeutic agent that targets human colon cancer stem cells," *Cell Death Dis*, vol. 6, p. e1801, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26136074

[118] H. T. Stuart, A. L. van Oosten, A. Radzisheuskaya, G. Martello, A. Miller, S. Dietmann, J. Nichols, and J. C. Silva, "Nanog amplifies stat3 activation and they synergistically induce the naive pluripotent program," *Curr Biol*, vol. 24, no. 3, pp. 340–6, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24462001

[119] H. Hernandez-Vargas, S. M. Rodriguez-Pinilla, M. Julian-Tendero, P. Sanchez-Rovira, C. Cuevas, A. Anton, M. J. Rios, J. Palacios, and G. Moreno-Bueno, "Gene expression profiling of breast cancer cells in response to gemcitabine: Nf-kappab pathway activation as a potential mechanism of resistance," *Breast Cancer Res Treat*, vol. 102, no. 2, pp. 157–72, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17039268

[120] H. Y. Lin, S. K. Hung, M. S. Lee, W. Y. Chiou, T. T. Huang, C. E. Tseng, L. Y. Shih, R. I. Lin, J. M. Lin, Y. H. Lai, C. B. Chang, F. C. Hsu, L. C. Chen, S. J. Tsai, Y. C. Su, S. C. Li, H. C. Lai, W. L. Hsu, D. W. Liu, C. K. Tai, S. F. Wu, and M. W. Chan, "Dna methylome analysis identifies epigenetic silencing of fhit as a determining factor for radiosensitivity in oral cancer: an outcome-predicting and treatment-implicating study," *Oncotarget*, vol. 6, no. 2, pp. 915–34, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25460508

[121] S. H. Kim, K. Joshi, R. Ezhilarasan, T. R. Myers, J. Siu, C. Gu, M. Nakano-Okuno, D. Taylor, M. Minata, E. P. Sulman, J. Lee, K. P. Bhat, A. E. Salcini, and I. Nakano, "Ezh2 protects glioma stem cells from radiation-induced cell death in a melk/foxm1-dependent manner," *Stem Cell Reports*, vol. 4, no. 2, pp. 226–38, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25601206

[122] Y. Du, L. Shi, T. Wang, Z. Liu, and Z. Wang, "Nanog sirna plus cisplatin may enhance the sensitivity of chemotherapy in esophageal cancer," *J Cancer Res Clin Oncol*, vol. 138, no. 10, pp. 1759–67, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22714588

[123] T. E. P. Consortium, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22955616

[124] A. Emad, J. Cairns, K. R. Kalari, L. Wang, and S. Sinha, "Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance," *Genome Biology*, vol. 18, no. 1, pp. 153–153, 2017. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1282-3

[125] E. J. Maier, B. C. Haynes, S. R. Gish, Z. A. Wang, M. L. Skowyra, A. L. Marulli, T. L. Doering, and M. R. Brent, "Model-driven mapping of transcriptional networks reveals the circuitry and dynamics of virulence regulation," *Genome Res*, vol. 25, no. 5, pp. 690–700, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25644834

[126] J. Qi and T. Michoel, "Context-specific transcriptional regulatory network inference from global gene expression maps using double two-way t-tests," *Bioinformatics*, vol. 28, no. 18, pp. 2325–32, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22962443

[127] G. Geeven, R. E. van Kesteren, A. B. Smit, and M. C. de Gunst, "Identification of context-specific gene regulatory networks with gemula–gene expression modeling using lasso," *Bioinformatics*, vol. 28, no. 2, pp. 214–21, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22106333

[128] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, "Generank: using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, p. 233, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16176585

[129] X. Chen, W. Jiang, Q. Wang, T. Huang, P. Wang, Y. Li, X. Chen, Y. Lv, and X. Li, "Systematically characterizing and prioritizing chemosensitivity related gene based on gene ontology and protein interaction network," *BMC Med Genomics*, vol. 5, p. 43, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23031817

[130] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. J. H. Penninx, R. Jansen, E. J. C. de Geus, D. I. Boomsma, F. A. Wright, P. F. Sullivan, E. Nikkola, M. Alvarez, M. Civelek, A. J. Lusis, T. Lehtimaki, E. Raitoharju, M. Kahonen, I. Seppala, O. T. Raitakari, J. Kuusisto, M. Laakso, A. L. Price, P. Pajukanta, and B. Pasaniuc, "Integrative approaches for large-scale transcriptome-wide association studies," *Nat Genet*, vol. 48, no. 3, pp. 245–252, 2016. [Online]. Available: http://dx.doi.org/10.1038/ng.3506

[131] A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang, A. L. Bracha, T. Liefeld, M. Wawer, J. C. Gilbert, A. J. Wilson, N. Stransky, G. V. Kryukov, V. Dancik, J. Barretina, L. A. Garraway, C. S. Hon, B. Munoz, J. A. Bittker, B. R. Stockwell, D. Khabele, A. M. Stern, P. A. Clemons, A. F. Shamji, and S. L. Schreiber, "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules," *Cell*, vol. 154, no. 5, pp. 1151–1161, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23993102

[132] N. Kawamura, K. Nimura, H. Nagano, S. Yamaguchi, N. Nonomura, and Y. Kaneda, "Crispr/cas9-mediated gene knockout of nanog and nanogp8 decreases the malignant potential of prostate cancer cells," *Oncotarget*, vol. 6, no. 26, pp. 22 361–74, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26087476

[133] C. R. Jeter, B. Liu, X. Liu, X. Chen, C. Liu, T. Calhoun-Davis, J. Repass, H. Zaehres, J. J. Shen, and D. G. Tang, "Nanog promotes cancer stem cell characteristics and prostate cancer resistance to androgen deprivation," *Oncogene*, vol. 30, no. 36, pp. 3833–45, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21499299

[134] P. Ordentlich, Y. Yan, S. Zhou, and R. A. Heyman, "Identification of the antineoplastic agent 6-mercaptopurine as an activator of the orphan nuclear hormone receptor nurr1," *J Biol Chem*, vol. 278, no. 27, pp. 24 791–9, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12709433

[135] K. Saijo, B. Winner, C. T. Carson, J. G. Collier, L. Boyer, M. G. Rosenfeld, F. H. Gage, and C. K. Glass, "A nurr1/corest pathway in microglia and astrocytes protects dopaminergic neurons from inflammation-induced death," *Cell*, vol. 137, no. 1, pp. 47–59, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19345186

[136] A. Nishimoto, N. Kugimiya, T. Hosoyama, T. Enoki, T. S. Li, and K. Hamano, "Hif-1alpha activation under glucose deprivation plays a central role in the acquisition of anti-apoptosis in human colon cancer cells," *Int J Oncol*, vol. 44, no. 6, pp. 2077–84, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24718784

[137] A. M. Ordelheide, F. Gerst, O. Rothfuss, M. Heni, C. Haas, I. Thielker, S. Herzberg-Schafer, A. Bohm, F. Machicao, S. Ullrich, N. Stefan, A. Fritsche, H. U. Haring, and H. Staiger, "Nor-1, a novel incretin-responsive regulator of insulin genes and insulin secretion," *Mol Metab*, vol. 2, no. 3, pp. 243–55, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24044104

[138] A. Chalastanis, V. Penard-Lacronique, M. Svrcek, V. Defaweux, N. Antoine, O. Buhard, S. Dumont, B. Fabiani, I. Renault, E. Tubacher, J. F. Flejou, H. Te Riele, A. Duval, and M. Muleris, "Azathioprine-induced carcinogenesis in mice according to msh2 genotype," *J Natl Cancer Inst*, vol. 102, no. 22, pp. 1731–40, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20923998

[139] S. Thota, A. D. Viny, H. Makishima, B. Spitzer, T. Radivoyevitch, B. Przychodzen, M. A. Sekeres, R. L. Levine, and J. P. Maciejewski, "Genetic alterations of the cohesin complex genes in myeloid malignancies," *Blood*, vol. 124, no. 11, pp. 1790–8, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25006131

[140] B. C. Valdez, G. Wang, D. Murray, Y. Nieto, Y. Li, J. Shah, F. Turturro, M. Wang, D. M. Weber, R. E. Champlin, M. H. Qazilbash, and B. S. Andersson, "Mechanistic studies on the synergistic cytotoxicity of the nucleoside analogs gemcitabine and clofarabine in multiple myeloma: relevance of p53 and its clinical implications," *Exp Hematol*, vol. 41, no. 8, pp. 719–30, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23648290

[141] L. M. Karnitz, K. S. Flatten, J. M. Wagner, D. Loegering, J. S. Hackbarth, S. J. H. Arlander, B. T. Vroman, M. B. Thomas, Y.-U. Baek, K. M. Hopkins, H. B. Lieberman, J. Chen, W. A. Cliby, and S. H. Kaufmann, "Gemcitabine-induced activation of checkpoint signaling pathways that affect tumor cell survival," *Molecular Pharmacology*, vol. 68, no. 6, pp. 1636–1644, 2005.

[142] E. Sanchez-Tillo, L. Fanlo, L. Siles, S. Montes-Moreno, A. Moros, G. Chiva-Blanch, R. Estruch, A. Martinez, D. Colomer, B. Gyorffy, G. Roue, and A. Postigo, "The emt activator zeb1 promotes tumor growth and determines differential response to chemotherapy in mantle cell lymphoma," *Cell Death Differ*, vol. 21, no. 2, pp. 247–57, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24013721

[143] S. Jagtap, K. Meganathan, J. Gaspar, V. Wagh, J. Winkler, J. Hescheler, and A. Sachinidis, "Cytosine arabinoside induces ectoderm and inhibits mesoderm expression in human embryonic stem cells during multilineage differentiation," *Br J Pharmacol*, vol. 162, no. 8, pp. 1743–56, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21198554

[144] B. Chen, J. Liu, Q. Chang, K. Beezhold, Y. Lu, and F. Chen, "Jnk and stat3 signaling pathways converge on akt-mediated phosphorylation of ezh2 in bronchial epithelial cells induced by arsenic," *Cell Cycle*, vol. 12, no. 1, pp. 112–21, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23255093

[145] J. Masliah-Planchon, I. Bieche, J. M. Guinebretiere, F. Bourdeaut, and O. Delattre, "Swi/snf chromatin remodeling and human malignancies," *Annu Rev Pathol*, vol. 10, pp. 145–71, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25387058

[146] H. H. Chen and M. T. Kuo, "Overcoming platinum drug resistance with copper-lowering agents," *Anticancer Res*, vol. 33, no. 10, pp. 4157–61, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24122978

[147] J. Y. Chun, Y. Hu, E. Pinder, J. Wu, F. Li, and A. C. Gao, "Selenium inhibition of survivin expression by preventing sp1 binding to its promoter," *Mol Cancer Ther*, vol. 6, no. 9, pp. 2572–80, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17876054

[148] K. J. Campbell, J. M. Witty, S. Rocha, and N. D. Perkins, "Cisplatin mimics arf tumor suppressor regulation of rela (p65) nuclear factor-kappab transactivation," *Cancer Res*, vol. 66, no. 2, pp. 929–35, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16424027

[149] A. Msaki, A. M. Sanchez, L. F. Koh, B. Barre, S. Rocha, N. D. Perkins, and R. F. Johnson, "The role of rela (p65) threonine 505 phosphorylation in the regulation of cell growth, survival, and migration," *Mol Biol Cell*, vol. 22, no. 17, pp. 3032–40, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21737676

[150] A. Kothandapani, K. Gopalakrishnan, B. Kahali, D. Reisman, and S. M. Patrick, "Downregulation of swi/snf chromatin remodeling factor subunits modulates cisplatin cytotoxicity," *Exp Cell Res*, vol. 318, no. 16, pp. 1973–86, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22721696

[151] C. A. Rabik and M. E. Dolan, "Molecular mechanisms of resistance and toxicity associated with platinating agents," *Cancer Treat Rev*, vol. 33, no. 1, pp. 9–23, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17084534

[152] C. Guo, J. Ding, L. Yao, L. Sun, T. Lin, Y. Song, L. Sun, and D. Fan, "Tumor suppressor gene runx3 sensitizes gastric cancer cells to chemotherapeutic drugs by downregulating bcl-2, mdr-1 and mrp-1," *Int J Cancer*, vol. 116, no. 1, pp. 155–60, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15756676

[153] N. Xu, C. Shen, Y. Luo, L. Xia, F. Xue, Q. Xia, and J. Zhang, "Upregulated mir-130a increases drug resistance by regulating runx3 and wnt signaling in cisplatin-treated hcc cell," *Biochem Biophys Res Commun*, vol. 425, no. 2, pp. 468–72, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22846564

[154] B. K. Biswal, M. J. Beyrouthy, M. P. Hever-Jardine, D. Armstrong, C. R. Tomlinson, B. C. Christensen, C. J. Marsit, and M. J. Spinella, "Acute hypersensitivity of pluripotent testicular cancer-derived embryonal carcinoma to low-dose 5-aza deoxycytidine is associated with global dna damage-associated p53 activation, anti-pluripotency and dna demethylation," *PLoS One*, vol. 7, no. 12, p. e53003, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23300844

[155] S. S. Dadarkar, L. C. Fonseca, A. D. Thakkar, P. B. Mishra, A. K. Rangasamy, and M. Padigaru, "Effect of nephrotoxicants and hepatotoxicants on gene expression profile in human peripheral blood mononuclear cells," *Biochem Biophys Res Commun*, vol. 401, no. 2, pp. 245–50, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20849824

[156] A. Ciccia, A. V. Nimonkar, Y. Hu, I. Hajdu, Y. J. Achar, L. Izhar, S. A. Petit, B. Adamson, J. C. Yoon, S. C. Kowalczykowski, D. M. Livingston, L. Haracska, and S. J. Elledge, "Polyubiquitinated pcna recruits the zranb3 translocase to maintain genomic integrity after replication stress," *Mol Cell*, vol. 47, no. 3, pp. 396–409, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22704558

[157] A. Yoshimura, M. Seki, T. Hayashi, Y. Kusa, S. Tada, Y. Ishii, and T. Enomoto, "Functional relationships between rad18 and wrnip1 in vertebrate cells," *Biol Pharm Bull*, vol. 29, no. 11, pp. 2192–6, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17077513

[158] W. K. Low, S. W. Kong, and M. G. Tan, "Ototoxicity from combined cisplatin and radiation treatment: an in vitro study," *Int J Otolaryngol*, vol. 2010, p. 523976, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21151649

[159] A. Limonciel, K. Moenks, S. Stanzel, G. L. Truisi, C. Parmentier, L. Aschauer, A. Wilmes, L. Richert, P. Hewitt, S. O. Mueller, A. Lukas, A. Kopp-Schneider, M. O. Leonard, and P. Jennings, "Transcriptomics hit the target: Monitoring of ligand-activated and stress response pathways for chemical testing," *Toxicol In Vitro*, vol. 30, no. 1 Pt A, pp. 7–18, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25596134

[160] Y. S. Lu, P. Y. Yeh, S. E. Chuang, M. Gao, M. L. Kuo, and A. L. Cheng, "Glucocorticoids enhance cytotoxicity of cisplatin via suppression of nf-kappab activation in the glucocorticoid receptor-rich human cervical carcinoma cell line siha," *J Endocrinol*, vol. 188, no. 2, pp. 311–9, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16461557

[161] T. C. Hour, Y. L. Lai, C. I. Kuan, C. K. Chou, J. M. Wang, H. Y. Tu, H. T. Hu, C. S. Lin, W. J. Wu, Y. S. Pu, E. Sterneck, and A. M. Huang, "Transcriptional up-regulation of sod1 by cebpd: a potential target for cisplatin resistant human urothelial carcinoma cells," *Biochem Pharmacol*, vol. 80, no. 3, pp. 325–34, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20385105

[162] S. Tardito, C. Isella, E. Medico, L. Marchio, E. Bevilacqua, M. Hatzoglou, O. Bussolati, and R. Franchi-Gazzola, "The thioxotriazole copper(ii) complex a0 induces endoplasmic reticulum stress and paraptotic death in human cancer cells," *J Biol Chem*, vol. 284, no. 36, pp. 24 306–19, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19561079

[163] B. Gyorffy, P. Surowiak, O. Kiesslich, C. Denkert, R. Schafer, M. Dietel, and H. Lage, "Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations," *Int J Cancer*, vol. 118, no. 7, pp. 1699–712, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16217747

[164] S. D. Vangipuram, S. A. Buck, and W. D. Lyman, "Wnt pathway activity confers chemoresistance to cancer stem-like cells in a neuroblastoma cell line," *Tumour Biol*, vol. 33, no. 6, pp. 2173–83, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22886526

[165] B. A. de Koning, D. J. Lindenbergh-Kortleve, R. Pieters, H. A. Buller, I. B. Renes, and A. W. Einerhand, "Alterations in epithelial and mesenchymal intestinal gene expression during doxorubicin-induced mucositis in mice," *Dig Dis Sci*, vol. 52, no. 8, pp. 1814–25, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17415656

[166] M. Bernard, E. Delabesse, S. Novault, O. Hermine, and E. A. Macintyre, "Antiapoptotic effect of ectopic tal1/scl expression in a human leukemic t-cell line," *Cancer Res*, vol. 58, no. 12, pp. 2680–7, 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9635597

[167] W. Lu, Z. Chen, H. Zhang, Y. Wang, Y. Luo, and P. Huang, "Znf143 transcription factor mediates cell survival through upregulation of the gpx1 activity in the mitochondrial respiratory dysfunction," *Cell Death Dis*, vol. 3, p. e422, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23152058

[168] Y. Ge, T. L. Jensen, D. A. Tatman, M. L. Stout, S. A. Buck, Y. Ravindranath, L. H. Matherly, and J. W. Taub, "Role of usf1 in the differential expression of the human deoxycytidine kinase gene in acute myeloid leukemia," *Leukemia*, vol. 19, no. 4, pp. 677–9, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15729384

[169] Y. Ge, T. L. Jensen, L. H. Matherly, and J. W. Taub, "Physical and functional interactions between usf and sp1 proteins regulate human deoxycytidine kinase promoter activity," *J Biol Chem*, vol. 278, no. 50, pp. 49 901–10, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14514691

[170] Y. He, S. Sun, H. Sha, Z. Liu, L. Yang, Z. Xue, H. Chen, and L. Qi, "Emerging roles for xbp1, a super transcription factor," *Gene Expr*, vol. 15, no. 1, pp. 13–25, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21061914

[171] L. Romero-Ramirez, H. Cao, D. Nelson, E. Hammond, A. H. Lee, H. Yoshida, K. Mori, L. H. Glimcher, N. C. Denko, A. J. Giaccia, Q. T. Le, and A. C. Koong, "Xbp1 is essential for survival under hypoxic conditions and is required for tumor growth," *Cancer Res*, vol. 64, no. 17, pp. 5943–7, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15342372

[172] Y. Kawano, Y. Kikukawa, S. Fujiwara, N. Wada, Y. Okuno, H. Mitsuya, and H. Hata, "Hypoxia reduces cd138 expression and induces an immature and stem cell-like transcriptional program in myeloma cells," *Int J Oncol*, vol. 43, no. 6, pp. 1809–16, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24126540

[173] J. Fiedler, K. Breckwoldt, C. W. Remmele, D. Hartmann, M. Dittrich, A. Pfanne, A. Just, K. Xiao, M. Kunz, T. Muller, A. Hansen, R. Geffers, T. Dandekar, T. Eschenhagen, and T. Thum, "Development of long noncoding rna-based strategies to modulate tissue vascularization," *J Am Coll Cardiol*, vol. 66, no. 18, pp. 2005–15, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26516004

[174] G. M. Kasof, L. Goyal, and E. White, "Btf, a novel death-promoting transcriptional repressor that interacts with bcl-2-related proteins," *Mol Cell Biol*, vol. 19, no. 6, pp. 4390–404, 1999. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10330179

[175] A. Pines, C. D. Kelstrup, M. G. Vrouwe, J. C. Puigvert, D. Typas, B. Misovic, A. de Groot, L. von Stechow, B. van de Water, E. H. Danen, H. Vrieling, L. H. Mullenders, and J. V. Olsen, "Global phosphoproteome profiling reveals unanticipated networks responsive to cisplatin treatment of embryonic stem cells," *Mol Cell Biol*, vol. 31, no. 24, pp. 4964–77, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22006019

[176] G. Hendriks, M. Atallah, B. Morolli, F. Calleja, N. Ras-Verloop, I. Huijskens, M. Raamsman, B. van de Water, and H. Vrieling, "The toxtracker assay: novel gfp reporter systems that provide mechanistic insight into the genotoxic properties of chemicals," *Toxicol Sci*, vol. 125, no. 1, pp. 285–98, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22003191

[177] A. Baldi, M. T. Piccolo, M. R. Boccellino, A. Donizetti, I. Cardillo, R. La Porta, L. Quagliuolo, E. P. Spugnini, F. Cordero, G. Citro, M. Menegozzo, R. A. Calogero, and S. Crispi, "Apoptosis induced by piroxicam plus cisplatin combined treatment is triggered by p21 in mesothelioma," *PLoS One*, vol. 6, no. 8, p. e23569, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21858171

[178] L. Lundholm, P. Haag, D. Zong, T. Juntti, B. Mork, R. Lewensohn, and K. Viktorsson, "Resistance to dna-damaging treatment in non-small cell lung cancer tumor-initiating cells involves reduced dna-pk/atm activation and diminished cell cycle arrest," *Cell Death Dis*, vol. 4, p. e478, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23370278

[179] A. B. Hall, D. Newsome, Y. Wang, D. M. Boucher, B. Eustace, Y. Gu, B. Hare, M. A. Johnson, S. Milton, C. E. Murphy, D. Takemoto, C. Tolman, M. Wood, P. Charlton, J. D. Charrier, B. Furey, J. Golec, P. M. Reaper, and J. R. Pollard, "Potentiation of tumor responses to dna damaging therapy by the selective atr inhibitor vx-970," *Oncotarget*, vol. 5, no. 14, pp. 5674–85, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25010037

[180] K. A. Lewis, K. K. Lilly, E. A. Reynolds, W. P. Sullivan, S. H. Kaufmann, and W. A. Cliby, "Ataxia telangiectasia and rad3-related kinase contributes to cell cycle arrest and survival after cisplatin but not oxaliplatin," *Mol Cancer Ther*, vol. 8, no. 4, pp. 855–63, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19372558

[181] A. T. Tubbs, Y. Dorsett, E. Chan, B. Helmink, B. S. Lee, P. Hung, R. George, A. L. Bredemeyer, A. Mittal, R. V. Pappu, D. Chowdhury, N. Mosammaparast, M. S. Krangel, and B. P. Sleckman, "Kap-1 promotes resection of broken dna ends not protected by gamma-h2ax and 53bp1 in g(1)-phase lymphocytes," *Mol Cell Biol*, vol. 34, no. 15, pp. 2811–21, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24842905

[182] S. J. Chiu, Y. J. Lee, T. S. Hsu, and W. S. Chen, "Oxaliplatin-induced gamma-h2ax activation via both p53-dependent and -independent pathways but is not associated with cell cycle arrest in human colorectal cancer cells," *Chem Biol Interact*, vol. 182, no. 2-3, pp. 173–82, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19735649

[183] M. Hu, X. Fu, Y. Cui, S. Xu, Y. Xu, Q. Dong, and L. Sun, "Expression of kap1 in epithelial ovarian cancer and its correlation with drug-resistance," *Int J Clin Exp Med*, vol. 8, no. 10, pp. 17 308–20, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26770323

[184] M. Davis, J. Li, E. Knight, S. R. Eldridge, K. K. Daniels, and P. R. Bushel, "Toxicogenomics profiling of bone marrow from rats treated with topotecan in combination with oxaliplatin: a mechanistic strategy to inform combination toxicity," *Front Genet*, vol. 6, p. 14, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25729387

[185] O. M. Alian, A. S. Azmi, and R. M. Mohammad, "Network insights on oxaliplatin anti-cancer mechanisms," *Clin Transl Med*, vol. 1, no. 1, p. 26, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23369220

[186] H. Tang, Y. J. Liu, M. Liu, and X. Li, "Establishment and gene analysis of an oxaliplatin-resistant colon cancer cell line thc8307/l-ohp," *Anticancer Drugs*, vol. 18, no. 6, pp. 633–9, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17762391

[187] U. Wellner, J. Schubert, U. C. Burk, O. Schmalhofer, F. Zhu, A. Sonntag, B. Waldvogel, C. Vannier, D. Darling, A. zur Hausen, V. G. Brunton, J. Morton, O. Sansom, J. Schuler, M. P. Stemmler, C. Herzberger, U. Hopt, T. Keck, S. Brabletz, and T. Brabletz, "The emt-activator zeb1 promotes tumorigenicity by repressing stemness-inhibiting micrornas," *Nat Cell Biol*, vol. 11, no. 12, pp. 1487–95, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19935649

[188] Y. Zhou, G. Wan, R. Spizzo, C. Ivan, R. Mathur, X. Hu, X. Ye, J. Lu, F. Fan, L. Xia, G. A. Calin, L. M. Ellis, and X. Lu, "mir-203 induces oxaliplatin resistance in colorectal cancer cells by negatively regulating atm kinase," *Mol Oncol*, vol. 8, no. 1, pp. 83–92, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24145123

[189] T. T. Vellinga, T. Borovski, V. C. de Boer, S. Fatrai, S. van Schelven, K. Trumpi, A. Verheem, N. Snoeren, B. L. Emmink, J. Koster, I. H. Rinkes, and O. Kranenburg, "Sirt1/pgc1alpha-dependent increase in oxidative phosphorylation supports chemotherapy resistance of colon cancer," *Clin Cancer Res*, vol. 21, no. 12, pp. 2870–9, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25779952

[190] S. Ray, Y. Lu, S. H. Kaufmann, W. C. Gustafson, J. E. Karp, I. Boldogh, A. P. Fields, and A. R. Brasier, "Genomic mechanisms of p210bcr-abl signaling: induction of heat shock protein 70 through the gata response element confers resistance to paclitaxel-induced apoptosis," *J Biol Chem*, vol. 279, no. 34, pp. 35 604–15, 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15155749

[191] M. V. Aguirre, J. S. Todaro, J. A. Juaristi, and N. C. Brandan, "Murine erythropoietic impairment induced by paclitaxel: interactions of gata-1 and erythroid kruppel-like transcription factors, apoptotic related proteins and erythropoietin receptor," *Eur J Pharmacol*, vol. 636, no. 1-3, pp. 42–51, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20353770

[192] R. Drago-Ferrante, A. Santulli, R. Di Fiore, M. Giuliano, G. Calvaruso, G. Tesoriere, and R. Vento, "Low doses of paclitaxel potently induce apoptosis in human retinoblastoma y79 cells by up-regulating e2f1," *Int J Oncol*, vol. 33, no. 4, pp. 677–87, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18813780

[193] A. J. Russo, P. G. Magro, Z. Hu, W. W. Li, R. Peters, J. Mandola, D. Banerjee, and J. R. Bertino, "E2f-1 overexpression in u2os cells increases cyclin b1 levels and cdc2 kinase activity and sensitizes cells to antimitotic agents," *Cancer Res*, vol. 66, no. 14, pp. 7253–60, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16849574

[194] L. Ma, L. Nie, J. Liu, B. Zhang, S. Song, M. Sun, J. Yang, Y. Yang, X. Fang, S. Hu, Y. Zhao, and J. Yu, "An rna-seq-based gene expression profiling of radiation-induced tumorigenic mammary epithelial cells," *Genomics Proteomics Bioinformatics*, vol. 10, no. 6, pp. 326–35, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23317700

[195] C. Shao, M. Folkard, K. D. Held, and K. M. Prise, "Estrogen enhanced cell-cell signalling in breast cancer cells exposed to targeted irradiation," *BMC Cancer*, vol. 8, p. 184, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18590532

[196] P. Salomoni, R. Bernardi, S. Bergmann, A. Changou, S. Tuttle, and P. P. Pandolfi, "The promyelocytic leukemia protein pml regulates c-jun function in response to dna damage," *Blood*, vol. 105, no. 9, pp. 3686–90, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15626733

[197] Z. X. Xu, A. Timanova-Atanasova, R. X. Zhao, and K. S. Chang, "Pml colocalizes with and stabilizes the dna damage response protein topbp1," *Mol Cell Biol*, vol. 23, no. 12, pp. 4247–56, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12773567

[198] J. Y. Chan, L. Li, Y. H. Fan, Z. M. Mu, W. W. Zhang, and K. S. Chang, "Cell-cycle regulation of dna damage-induced expression of the suppressor gene pml," *Biochem Biophys Res Commun*, vol. 240, no. 3, pp. 640–6, 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9398618

[199] Y. Han, Y. Zhang, L. H. Yang, X. Y. Mi, S. D. Dai, Q. C. Li, H. T. Xu, J. H. Yu, G. Li, J. Zhao, C. Han, X. M. Yuan, and E. H. Wang, "X-radiation inhibits histone deacetylase 1 and 2, upregulates axin expression and induces apoptosis in non-small cell lung cancer," *Radiat Oncol*, vol. 7, p. 183, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23110995

[200] M. New, H. Olzscha, and N. B. La Thangue, "Hdac inhibitor-based therapies: can we interpret the code?" *Mol Oncol*, vol. 6, no. 6, pp. 637–56, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23141799

[201] J. A. Fielhaber, Y. S. Han, J. Tan, S. Xing, C. M. Biggs, K. B. Joung, and A. S. Kristof, "Inactivation of mammalian target of rapamycin increases stat1 nuclear content and transcriptional activity in alpha4- and protein phosphatase 2a-dependent fashion," *J Biol Chem*, vol. 284, no. 36, pp. 24 341–53, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19553685

[202] X. Liu, X. Wang, Y. Bi, P. Bu, and M. Zhang, "The histone demethylase phf8 represses cardiac hypertrophy upon pressure overload," *Exp Cell Res*, vol. 335, no. 1, pp. 123–34, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25921086

[203] A. Kreisler, P. L. Strissel, R. Strick, S. B. Neumann, U. Schumacher, and C. M. Becker, "Regulation of the nrsf/rest gene by methylation and creb affects the cellular phenotype of small-cell lung cancer," *Oncogene*, vol. 29, no. 43, pp. 5828–38, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20697351

[204] J. S. Kim, B. S. Kim, J. Kim, C. S. Park, and I. Y. Chung, "The phosphoinositide-3-kinase/akt pathway mediates the transient increase in nanog expression during differentiation of f9 cells," *Arch Pharm Res*, vol. 33, no. 7, pp. 1117–25, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20661723

[205] B. Chen, Z. Xue, G. Yang, B. Shi, B. Yang, Y. Yan, X. Wang, D. Han, Y. Huang, and W. Dong, "Akt-signal integration is involved in the differentiation of embryonal carcinoma cells," *PLoS One*, vol. 8, no. 6, p. e64877, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23762260

[206] D. J. Weisenberger, "Characterizing dna methylation alterations from the cancer genome atlas," *J Clin Invest*, vol. 124, no. 1, pp. 17–23, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24382385

[207] C. Houillier, X. Wang, G. Kaloshi, K. Mokhtari, R. Guillevin, J. Laffaire, S. Paris, B. Boisselier, A. Idbaih, F. Laigle-Donadey, K. Hoang-Xuan, M. Sanson, and J. Y. Delattre, "Idh1 or idh2 mutations predict longer survival and response to temozolomide in low-grade gliomas," *Neurology*, vol. 75, no. 17, pp. 1560–6, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20975057

[208] P. Guilhamon, M. Eskandarpour, D. Halai, G. A. Wilson, A. Feber, A. E. Teschendorff, V. Gomez, A. Hergovich, R. Tirabosco, M. Fernanda Amary, D. Baumhoer, G. Jundt, M. T. Ross, A. M. Flanagan, and S. Beck, "Meta-analysis of idh-mutant cancers identifies ebf1 as an interaction partner for tet2," *Nat Commun*, vol. 4, p. 2166, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23863747